Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Infer - A Full Scale B2B Sales Predictor

Author: Kailainathan Muthiah Kasinathan (2662871)

1st supervisor:	Dr. Adam S.Z. Belloum
daily supervisor:	Dr. Adam S.Z. Belloum
2nd reader:	supervisor name

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

August 17, 2021

"Sometimes, the wrong train takes you to the right station."
from
Crash Landing on You

Abstract

Sales forecasting is an important step in a Business to Business (B2B) company's financial planning. Sales forecasting helps to estimate the future sales and is mostly based on historical data. Accurate estimation of sales allows a company to find out about the cash movement and also about the variation in demand of the products. The area of Machine Learning provides a way to analyse the historic data from the sales pipeline of the company. Keeping Sales Forecast as the domain, The thesis aims to achieve three goals - Predict whether a sales opportunity will be won or lost, formulate a methodology to get the aggregated forecast values at an Area level from the predicted output and to finally develop visualizations to monitor sales pipeline data's hygiene and health aspects. This is aimed to provide an alternative to conventional forecasting process and also to see how data driven forecasting can be used to challenge the conventional process. CatBoost Classifier predicts the sales opportunities and classifies them as won or lost, and a mathematical approach is used to adjust the aggregated forecast value. Tableau is used to develop dashboards to monitor the sales pipeline's hygiene and health. We obtained acceptable results from this approach when compared to forecasts formulated by the conventional process. This thesis also exposed us to a variety of challenges when we do B2B forecasting with machine learning, which are explained as part of the thesis.

Acknowledgements

I would like to thank Dr.Adam Belloum for assisting me throughout the entire scope of the thesis and also for guiding me in the right direction. I would like to thank my intern manager at NetApp, Mr.Minh HoDac and Senior financial Analyst, Mr.Németh Mate for introducing me into this research area and helping out with the data understanding. I would like to thank all the members from the sales and finance teams of NetApp, who were kind enough to help with answers to my questions related to the thesis. I would also like to thank Mr. Rajesh Gupta for helping me out in brainstorming ideas and taking part in discussions which helped improve the thesis. On a final note, I would like to thank my family for everything.Special thanks to Mr.Asif Hassan, Ms.Diksha Das and Mr.Navin Kumar, who were always a constant support during the entire duration of my masters.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Glossary

Account A success full Opportunity which generates revenue for the company.

Actual Sales A list of all completed opportunities which turned into revenue for the company.

B2B Sales Business to Business. Selling products between two Businesses.

B2C Sales Business to Consumer. Selling products between a business and a consumer mostly an individual.

CRM Customer Relationship Management tool.Tool to administer interactions with customers.

Lead A customer willing to buy from a company

Opportunity Also referred to as Deal. Provides details about what the customer intends to buy or simply Deal details.

Sales Health Health is good if there are more potential opportunities present in the sales pipeline that will be completed in the future and vice versa

Sales Hygiene Ensuring the quality of the data in the Sales Pipeline.

Sales Pipeline A list of all expected/completed Opportunities.

LIST OF TABLES

1

Introduction

Sales plays a key role in determining the success of an Organization.In layman terms, Sales is selling a product or a service to a customer in return for a remuneration.Business to Business (B2B) Sales is selling directly to other Businesses, whereas Business to Customer (B2C) Sales is selling directly to an individual customer. Organizations can be classified into these two categories based on how they sell their products or solutions. To understand B2B and B2C sales, we provide an example for each of these types. For B2B Sales, consider a Laptop. Suppose a display component is manufactured by a company "A" which is used by the Laptop manufactured by company "B" then A is said to be having a B2B Sales transaction since the sales is between two businesses. On the other hand, e-commerce sites are a very good example for B2C transactions as the sales is between a business and an individual. We only consider Business to Business(B2B) sales in accordance with the thesis. Businesses are often run based on Planning, and It's essential to understand how sales will be in the upcoming year to plan accordingly. This makes forecasting Sales an important component in the Financial Planning process.

Sales forecasting is the process of estimating the future sales based on the current and historical sales transactions. Why is Sales Forecasting important? Sales forecasting is an important jigsaw to the overall puzzle of Financial Planning. The results of Sales Forecasting essentially helps to understand the demand forecasting, cash Flows in an Organization and also helps maintain a better inventory. This relationship of Sales Forecasting also makes it a tough procedure since it needs to be highly accurate to make the dependent processes obtain high quality results. Accuracy of forecast is the primary metric to evaluate the quality of the forecast. An overestimated forecast can lead to higher expectations in sales and also lead to a high inventory, which can cause potential loss to the company.

Similarly, an underestimated forecast can cause the company serious troubles when there is a high demand and can also lead to a low inventory.

In a B2B setup, Sales Forecasting differs from the normal ways in which one forecasts sales of a supermarket or a toy store. Everything is based on Opportunities in the B2B setup, and it's easily stored and monitored using the ever improving Customer Relationship Management (CRM) systems. Opportunities are deals which are agreed in principle between the two parties and will be completed in a given time frame.B2B companies generally have an Opportunity Pipeline which will contain all these deals and their details. Traditionally, Sales teams just pick out opportunities from the Pipeline as per their knowledge or experience with the customer and use them to formulate the forecast. Even though this process works, It still does not provide enough evidence to support why a certain Opportunity should be picked or avoided in the forecast process. It's purely based on the individual's inputs, making this process unidirectional and biased based on the individual handling the opportunity. Sometimes this leads to underestimated or overestimated forecast.

With the rise of Machine learning, Systems are becoming intelligent in the way they understand the data. Machine learning opens up an interesting possibility for a system to learn from the data at hand, understand and predict the way the future looks. Forecasting is an interesting area of Application for Machine learning. Although machine learning is widely used in weather forecasting, forecasting the sales using historical selling data etc., It's still not being put in practice in terms of Sales Forecasting in B2B processes where we need to understand the opportunities to predict how the forecast will look like. This thesis aims to solve the problem at hand by building a predictor model that will classify the Opportunities into the two areas, whether it will be completed or not completed. We also try to identify and list out the best Machine learning algorithms which can be used in terms of commercially acceptable accuracy levels. The thesis was carried out at the Worldwide Financial Planning and Analysis team of NetApp - a hybrid cloud data services and data management company. The company's Sales data was used to model the predictor and to evaluate the results of the model being built. Upon completion, the model was deployed to be used as the Forecast predictor for the Worldwide Financial Planning and Analysis team.

This thesis is organized into the following chapters:

- Chapter 2 gives the Background where we discuss in detail about the opportunity cycle lifeline, Traditional approaches etc, and it concludes with drafting out our

research questions and their motivation.

- Chapter 3 explains the related work associated with machine learning in the Sales Forecasting domain.
- Chapter 4 explains the overview about the data, feature engineering approaches and feature selection approaches considered in the thesis.
- Chapter 5 explains the hybrid approach used in the thesis where we discuss the three components of the thesis namely - Machine learning algorithm, Mathematical approach and the visualizations. This chapter also explains the architecture of the model - Training flow and the active flow.
- Chapter 6 outlines the experiment setup and the evaluation of the results of the three components.
- Chapter 7 discusses the challenges faced by applying machine learning in a B2B sales forecasting setup.
- Chapter 8 briefs on the limitations and the future work associated with the thesis.
- We conclude the thesis report with Chapter 9 which explains the conclusions we obtained from performing this thesis.

2

Background

In this chapter, We explain in detail about the Opportunity Life Cycle to provide a better understanding of B2B Sales process. Additionally, we shed some light on the traditional way of formulating forecasts by the Sales Representatives. This chapter concludes by giving the aforementioned explanations and also lists out our research questions which will form the basis for the rest of the paper.

2.1 Understanding B2B Sales Process

With the advancement in technologies, Customer Relationship Management (CRM) Systems have come to the fore of the Sales processes. According to Shoemaker, CRM is a tool that blends all the essential information systems like Sales, Marketing and Service to build partnerships with customers. A recent survey by BuyerZone stated that 91 % of the companies with more than eleven employees now use CRM. CRM's have been adopted by almost all B2B businesses due to the benefits obtained from them stated by Peterson Et Al. like High efficiency, increased productivity of sales processes, better forecasting and performance and ability to document customer needs better.Although CRM is a wider topic area and many research activities have been carried out to identify their benefits, Usage, Implementation, effectiveness etc, we do not deep dive into that. We just establish the definition of CRM to understand the context of the Sales process in the following sections of this chapter.

2.1.1 B2B Sales Process - Stages

The B2B Sales process is similar to the Sales Funnel, and it involves the working of Sales, Marketing and Finance teams. It follows the seven stages involved in the sales funnel,

which can be broadly placed into three groups, namely -

- Lead Generation
- Opportunity Management
- Completion of Sales

The Lead generation process is the primary step. In this step, Marketing teams reach out to potential customers through different means such as Business meetings, E-mails etc. Once they establish a contact with the customer, they interact with the customers and explain about the products and how it can solve their problems. Upon interest from the customer to buy the product, the team hands over the details of the customer to the Sales teams, thereby marking the customer as a Lead.

Next comes the Opportunity management step. Once the Sales team takes over the handling of the lead, they start understanding the complex asks of the customer and device a quote. This quote is an estimate of the Sale of Products to the customer. The negotiations take place between the Sales team - ideally a Sales Representative and the customer about the quote and the solution offered to them. When both the parties come to a consensus, the deal gets flagged as an opportunity and goes into the Sales pipeline of the company. This opportunity then goes through different stages to reach the Won/Lost stage. The Won/Lost stage is essentially the final step in Opportunity Management post which it moves into the final step in the Sales process.When an opportunity is closed successfully, the customer essentially becomes an Account with the company. In the completion of Sales, both the Sales team and Finance teams play a role. As the opportunity gets completed, It contributes to the revenue of the company where the Finance teams start performing their activities like estimation of revenues, planning for the next Financial year etc. The Sales team also follows up on the order by keeping in touch with the customer for future orders, renewal of product subscriptions, reconfiguring the existing product etc. Where does CRM come into play? CRM contains all the information about the lead, Opportunity and Account and acts as the repository for the Sales information. Lead details are handled by the marketing team whereas the Sales teams constantly update the details about the opportunity such as sales stage, value changes in the opportunity, close date etc. This makes CRM a very handy tool from which a plethora of data can be obtained to perform analysis and also use them to build advanced models and automation's. For the context of the thesis, We look only into the Opportunity Management section of the Sales B2B process and skip the remainder.

Sales Stage	Forecast Category
Prospecting	Commit
Qualification	Best Case
Proposal	Pipeline
Acceptance	Omitted
Negotiation	Closed
Won	
Closed/Lost	

Figure 2.1: Different Stages and Forecast Categories

2.1.2 Opportunity Management

A brief insight into Opportunity Management was given in Subsection 2.1.1. In this part, we try to provide a detailed explanation into the different stages an Opportunity goes through before getting classified as Won or Lost Opportunity. We also outline other details associated to Opportunities in a Sales Pipeline.

Every Opportunity has to go through different Sales Stages and Forecast Categories when it resides in the Sales Pipeline. In general, Sales Stages act as an Opportunity tracker. Sales representatives constantly update this as the Opportunity progresses towards its final stage.Forecast category on the other hand helps to understand the probability that an Opportunity will get completed in the agreed time period. This is also updated by the Sales representatives. The Sales Representatives who own the Opportunity are called Opportunity owner, and they own and update the Opportunity details. Every company has its own way of determining the Sales stages. To maintain the consistency, We will look at the Sales stages and Forecast categories determined at the Company in which the thesis was carried out.

2.1.2.1 Sales Stages

Figure 2.1 shows the different types of Sales Stages. Prospecting stage is the one at which a lead or genuine piece of business is reviewed by the Sales team before moving into the Qualification stage. An opportunity is set to reach the Qualification stage when the team meets with the customer and has qualified that there is a business requirement and a potential budget to solve the need. Post the qualification of the business requirement,a solution is formulated as per the proposed budget and discussed with the customer.This contributes to the Proposal stage. Acceptance stage is where all information, proposals

and solution documentation with the customer, budget approvals and business acceptance take place for an Opportunity. Negotiation phase is where the customer engages with the sales team raising objections, engaging in next steps etc in order to make progress with the deal. This is also the phase in which the customer has agreed to terms and is finalizing the paperwork to invest in the solution provided. Won stage is where the Opportunity has been successfully completed and contributes to the revenue of the company. Closed/Lost is the stage where the deal fails to materialize. Opportunities are not mandated to go through all these sales stages.It depends on how the deal is progressing and in some stages it can directly go into the Won phase or Closed/Lost phase or can skip some stages too.

2.1.2.2 Forecast Category

As already mentioned in Subsection 2.1.2, Forecast Category is the probability or the likelihood of an Opportunity to complete in an agreed upon time period. The five different categories are shown in figure 2.1. Best Case means an Opportunity has high chances of getting completed. Committed means an Opportunity will be completed in the given time period. Closed means that the Opportunity has already been completed. Omitted refers to Opportunities that gets cancelled, and these deals go out of the sales Pipeline. Pipeline stage is the one at which the Opportunity is present but will not be completed in that agreed time period but will eventually be completed soon. In other words, the deal is delayed but not slipped from the Company's hands.

2.2 Conventional Forecasting Process

The Traditional Forecast process followed in the majority of B2B companies, including the company where this thesis is carried out, revolves around the Sales team. What contributes to the Forecast? As mentioned earlier in Chapter 1, the Opportunities in the Sales Pipeline contribute to the Forecast process. Many companies have different hierarchies as per which they formulate the forecast. Let's consider the procedure followed in the company where this thesis was carried out. Sales forecast is formulated by different personnel at different levels of Sales Territory Hierarchy. The Sales Territory is split into the following ways - District, Region, Multi Region, Area, Multi Area and Geography. For example, the Sales Representatives and District Sales Managers fall into the District level. At each level, with an increase in Sales Territory Hierarchy, there is an increase in the Seniority level of personnel.

The forecast process starts at the District level where the Sales representatives mark the deals which can be included as part of the forecast and submits it to the District manager who formulates the district level of forecast. It moves along the Sales hierarchy up till the Sales Multi Area level where it is finally submitted as the official forecast. At each level, the forecast can be modified based on the person's assumptions and knowledge. Also, the final forecast amount can be adjusted or overridden in order to account for risk or to add any excess opportunities. This finalized forecast which is purely driven based on the knowledge provided by individuals by their interactions with the customer, previous experiences with the customer, knowledge about the existing situation etc is submitted to be the official forecast. The forecast process is carried out every month by the finance team based on the input from the Sales team.

With this background about B2B Sales Forecast process, let's look at the problem and what is our research area, which will set the tone for the remaining part of the thesis.

2.3 Research Problem and Questions

The following points can be observed from the details provided in chapter 2.

- Pipeline is an important part of the Sales process, and it is essential that a company has a healthy pipeline in order to showcase and maintain a strong presence in the market.
- The conventional forecast process is purely driven by Individual's knowledge and assumption about the Opportunity.
- Ability to understand Opportunity conversion rate is minimal, i.e, Will an Opportunity get completed or not?

In theory, the thesis aims to solve an organizational challenge at the company and also aims to provide additional explanation towards the application of Machine Learning in B2B Sales Forecasting process. The following subsections provide the research questions this thesis will answer and the grounds on which it was considered.

2.3.0.1 Research Question - 1

Firstly, We try to solve the organizational challenge faced by the company - NetApp. The company's Financial Planning and Analysis team formulates the sales Forecasts depending on the inputs provided by the Sales team. Although this forecast is of acceptable quality,

there is a concern that the forecasts are purely driven by the Individual's knowledge. This constitutes to some commonly occurring problems with respect to Sales team and their ability to pick deals to be included as part of the forecast process. Some key problems are:

- If the sales team achieve their sales quota for the month, they can push Opportunities that have high likelihood to get completed to the next forecast period. This means some deals that should be a part of the forecast essentially misses out.
- The ability to mark the deals as part of the forecast process is based on individuals. This can lead to Human bias, repetitive behaviour and other cognitive biases.
- There is no means to challenge the Sales representatives on the deals they selected. It basically becomes a unidirectional process.
- Finally, not all factors are considered when selecting the deals. Limited knowledge is another area of concern, provided the capacity of data available.

This paved the way to look at Machine Learning methods to learn from the Pipeline data and design a predictor which predicts the possibility of the win/loss criteria for each opportunity and also helps in formulating the forecast. This way we can make a data driven approach and also look at an array of features available in the data and pick the ones which are more suitable to solve the problem at hand. Also, new products which are introduced never make it to the forecast in the current year. The forecast value is set to null, as there are no deals in the pipeline for the new product. A survey is carried out to find the expected reach of the product, but it does not get translated into the forecast.This paved the way for the first research question and its sub research questions, which is formulated as :-

1. Can a Sales Forecast formulated using Machine Learning challenge the Sales Forecast generated by the Knowledge of the sales team in terms of Quality?
    - Which Machine Learning Algorithm performs better with respect to the prediction?
    - How expert's knowledge can be incorporated into modelling the data?
    - How the model results can be evaluated?
    - What level of accuracy is accepted for the model to be deployed commercially?

2.3.0.2 Research Question - 2

Machine Learning models tend to need constant monitoring to ensure the predictions don't deteriorate over time. In other words, Production models suffer from the concept of Data drift. Data drift is defined as the change in input data compared to that of the training data, which causes the reduced the performance of the model. B2B Sales data has a bigger scope for constantly changing. There are also higher chances for unseen data to flow in as input data for the model. We formulate the following research question to understand the impact of Data drift and also the challenges which arise due to it in making the predictor model production ready.

1. How Data Drift plays a role in B2B Sales Data? Does it have a significant impact, and What can be done to minimize the changes in model performance, if any?

2.3.0.3 Research Question - 3

Building the predictor gives the ability to formulate the forecast procedures, but there is more to Pipeline than just forecasting. Understanding the count of Open Opportunities, count of deals to complete, and various other counts are provided by the tools available in CRM. But the ability to provide advanced analytics that drive the decision-making processes is still a key area for us to research. With the predictor output and the pipeline data, identifying the Key performance indicators (KPI's) and explaining them using Visualizations is needed which leads us to the final research question which is as follows:

1. What are the KPI's with respect to Sales Pipeline/Predictor Output and does using Visualizations a good way to understand the Sales Pipeline?
    - What kind of Visualizations can be used for this process?

3

Related Work

As already stated in the beginning, B2B Sales Forecasting takes up a bigger space in the process of Financial Planning and Analysis, and we try to apply machine learning to perform the process of forecasting. The application of Machine learning in B2C forecasting procedures is higher compared to its usage in formulating B2B forecasting procedures. In general, B2B sales data is quite complicated compared to B2C and the traditional approach of formulating the forecasts as discussed in section 2.2 is favoured more over any alternatives. This chapter gives an overview of experiments and studies carried out to understand and convey the use of machine learning in B2B sales forecasting procedures.

3.1 Machine Learning and B2B Forecasting

B2B Sales forecasting has a limited literature when it comes to the application of machine learning techniques in this domain. This is largely due to the fact that most of the organizations have not yet started exploring the usage of Artificial intelligence in the B2B domain, and also the reluctance to navigate away from the conventional method of forecasting. With that said, we will start looking into the different approaches and problem areas identified by the previous researchers.

Marko Bohanec et al. (2015) proposed a novel model based on machine learning techniques to counter the Sales forecasting problem.(1) They formulate a framework which will involve a classification algorithm at the heart of it and with this try to predict the sales opportunities. The output of the model is aimed to reduce the forecast error and also to provide insights about the model's prediction, which in turn triggers an organizational learning. Thereby, the authors try to create a feedback loop which will involve the machine learning model and via the insights trigger organizational learning. The input data to the

research was anonymized from the funding company and had only fewer instances of the data. To generate more instances, "semiArtificial" package of R was used to use the data for modelling purposes. RandomForest classifier was observed to be the high performing classifier and recorded accuracies of 96% and the model's output was visualized to form the feedback loop. Although this gave the idea that the machine learning can be applied in B2B sales forecasting domain, yet there were certain limitations. The limitation of this research was it had only limited training instances and much of them were created artificially, which possibly explains the high levels of classification accuracy obtained from this approach. The visualizations used to explain the models are complex which is hard to understand for a person coming from a sales background, which will reduce the participation of people in organizational learning.

Junchi Yan et al. (2015) proposed a model which will build a predictor but will look at things from a regression point of view. The model was aimed to make a predictor which will give a win propensity for each lead.(2) The methodology of this work is based on two steps - First being the collection of historical leads data and also the profile of the leads and the labels associated with the leads. The authors try to explain the synergy between the sales and marketing team and are more focused in determining the propensity score, which will potentially convey whether the lead will be won or not. Stephen Mortensen et al.(2019) decided to create an initial baseline for understanding the sales outcome propensity prediction in B2B setup.(3) They experimented with different classification algorithms like decision tree (4), RandomForest (5) and XGBoost (6) to come up with a good classifier and carried out the implementation using data from a Fortune 500 paper and packaging company. The main purpose of the research in addition to creating a baseline was to find out what features really contributed towards determining the win propensity. Again, RandomForest was found to be performing well with 82% accuracy and precision and recall were around the 77%-79% mark. Also, the authors experimented by providing individual RandomForest models for each business unit of the company and observed that the predictions were better since it operated based on the knowledge within the unit and also the execution time improved.

Brendan Duncan and Charles Elkan (2015) applied machine learning techniques to the B2B sales domain. The goal of the paper is different from predicting the outcome of the sales opportunities. Instead, the authors, try to determine the step before in the sales funnel. They build scoring two scoring models - namely Direct Qualification model and Full, Funnel model (7) which estimates the chance of a lead to get converted into an opportunity. This is more to understand whether the lead generated by the marketing

team will be converted into an opportunity, which potentially goes into the won or lost stage to result in the revenue generation process for the company. The research helps to potentially replace the human based lead scoring methodologies used in companies to understand whether a lead will get converted into opportunity or not. Although it is different to the context of the thesis, this still provides a valuable insight into how machine learning is used in the B2B sales forecasting area.

Alirezza Rezazadeh (2020) built an azure based machine learning approach to predict the outcome of sales opportunities. The paper proposes an end to end cloud based workflow to forecast the outcome (8). It takes the binary classification approach and uses a voting based classifier which is built with XGBoost and LightGBM (9) models. The research uses a B2B consulting firm's data belonging to different sectors such as healthcare, energy and finance. The implementation is done using azure and the framework consists of two pipelines - ML and Prediction pipeline which is used for training and for predicting the outcome on the live data. The author tries to formulate the decision boundaries for each of the different sectors and sets them as decision threshold which help to improve the classification accuracy for the individual sectors. The feature generation was another interesting aspect of this research, as the author creates lookup tables containing statistics for the categorical variables and merges them back to the original dataset at hand to create new features. The evaluation of the model's prediction was also slightly different. In addition to the classification metrics, the author makes use of the user generated probability score for each opportunity and compares them to the probabilities generated by the model to evaluate the outcome.

A forecast framework consisting of different types of machine learning models that would work based on different sales patterns or data changes was proposed by Xin Xu et al. in 2017. At the heart of the forecast engine lies different models namely - Probability models, Opportunity score aggregation models, time series models, Hybrid models and Neural network models and the optimal model is selected based on the request generated by the user(10). The probability models work to provide an aggregate level of forecast by grouping opportunities into time periods or groups preferred by the user, whereas the opportunity score aggregation models look into the history of the opportunity and calculate the likelihood that it can be closed within the time period specified. The time series also works on the aggregated level with respect to date/time/year. This methodology produced higher forecast accuracy levels, but the forecast and predictions happen at an aggregate level and not at an opportunity level. Timo Thiess et al. (2020) designed a predictor to find out about the B2B sales outcome of the opportunities as part of a project with MAN Energy Solutions (11). The research focuses on building a predictor model, but intends

to extend the explainability of the machine learning techniques in the B2B sales area. LightGBM classifier is used to classify the outcome of the opportunities and SHAP (12) is used to explain the machine learning model. The author also formulates certain steps which are termed as the design principles upon which their framework was built. The author with the help of SHAP and an interface explains the model's predictions for each opportunity with the intention to improve organizational learning.

With the brief look into the literature, we were able to find some common limitations which we try to overcome as part of our thesis in addition to solving the organizational challenge at hand. Research in this area often used limited training data or semi artificial training data, which eradicated many real life transformations happening in the nature of the Sales pipeline data at an organization. This might also influence the accuracy related to the classifications. Next, we were able to find that the some researches mentioned above, formulated the forecast value, but they didn't talk about the quality of the data except for one. For example, If the quality of the data is low on the sales value of the opportunities , then it will lead to a classifier that is performing exceptional at classifying the opportunity's outcome but not so much when it comes to getting the forecast value. Additionally, the feedback learning is a good approach to improve the model's performance and the thinking of the teams in context to improving the way they look at the outcomes, but it doesn't shed any light on the status of the sales pipeline and the quality of the data in the sales pipeline which can also be improved with the help of visualizations. So with all this and even some smaller limitations observed in mind, we formulate our modelling approach in the following chapters.

4

Data Overview

This chapter deals with the details about the B2B Sales dataset which was used. We try to initially provide an explanation about the data sources and then more into how it was transformed from its raw format into a processed input which could be fed to the machine learning pipeline to generate predictions. In this chapter, we also provide insights into how feature engineering and encoding methodologies were used and whether they yielded any results or not.

4.1 Data Explanation

B2B Sales dataset consists of data related to the Sales Pipeline. Our B2B Sales data is provided to us by NetApp. The Sales representatives enter quote details about prospective deals into the Quote interface system developed by a third party. So the dataset basically consists of upcoming, completed, missed and lost opportunities along with opportunity specific details. The data is stored in a CRM system which houses all levels of details associated to the data - Personal, Sale details, Product details, Customer details etc. We pull the data from the CRM and store it into a relational database and query this table to get the training data. This arrangement of connecting them to a relational table also has a reason behind it. For the machine learning aspect of the thesis, which will be explained in the later sections, We need to have a training pipeline and a production pipeline that influences us to make this decision of pulling the data from CRM into a relational database. As mentioned in various (reference papers), the sales data is noisy. The main driving factor behind this is that the values are updated by the sales people, which brings in the possibility of introducing human errors like abnormal sales value, incorrect updates to critical fields and ability to leave fields unfilled. The B2B Sales data collection is a purely human driven

approach, and it's not something like sensor data where the devices collect the data and the margin of error is lower.

The raw dataset obtained from the CRM consists of 179 features pertaining to different levels of information related to an opportunity. Although there were many dimensions to the dataset, We didn't make use of most of them as they did not provide any valuable insight to our goal. Upon discussion with the responsible people, we trimmed down the dataset to contain only 34 features, including both dimensions and measures.The reasoning behind the trimming down of the dataset is provided below:

- Most of the features were duplicate features, and they were marked redundant. This was due to the pending clean up of the CRM data.
- Many features reflected Opportunity value in different currencies, as the dataset contained worldwide data. As the forecasts were formulated in US Dollars, only values in USD was considered and the rest were removed.
- Drill down into product level details, areas, customer segmentation types were also present which were excluded.

Description	Value
Dataset Size	0.3M
Total Features	179
Selected Features	34

Table 4.1: Dataset - Size/Shape

4.2 Train and Test Data

Every machine learning algorithm needs data to produce results. General practice is to split the dataset into Training and test datasets to train and evaluate the model's performance. In our scenario, we have the historical pipeline data from the First quarter of 2018 to the third quarter of 2021 which was used as the training data and the fourth quarter of 2021 was used as the test dataset.

Description	Value	Duration of Data
Training Dataset	215.6K	2018Q1-2021Q3
Test Dataset	92.4K	2021Q4

Table 4.2: Train/Test - Details

4.2.1 Train Data Profiling

The training data is profiled to understand its composition. The profiling is explained in short below:

- Overview :- The training data consists of 215.6K records as stated in the previous section with 34 features. The dataset possesses a label which conveys whether an opportunity was won or lost. We have zero duplicate records and the missing cells % i.e, the number of cells found to have no data in the overall dataset seems to be 5.5 %. The dataset takes up a memory size of 94.4 MiB in total.
- Categorical Features :- Sales datasets always have higher number of categorical features as organizations use them to house the Sales territory details such as country, region etc and also details such as product type, Customer type too. In the training dataset, we have 29 Categorical features - Seven of them are highly cardinal.
- Figure 4.1 below shows the amount of missing data present in the dataset. The figure shows the count of values present in the dataset, split by different features. Figure 4.2 reflects the heatmap displaying the correlation between the variables.

4.3 Feature Engineering

This section explains in detail the preprocessing steps, encoding methodologies and the outlook of the processed input data.

4.3.1 Data Preprocessing

In the Data Preprocessing stage, the aim is to prepare the raw data into the desired format and clean the dataset. The following steps were carried out as part of the preprocessing stage:

- Firstly, Unknown values in features representing Sales territory and Sales Team details were removed to.

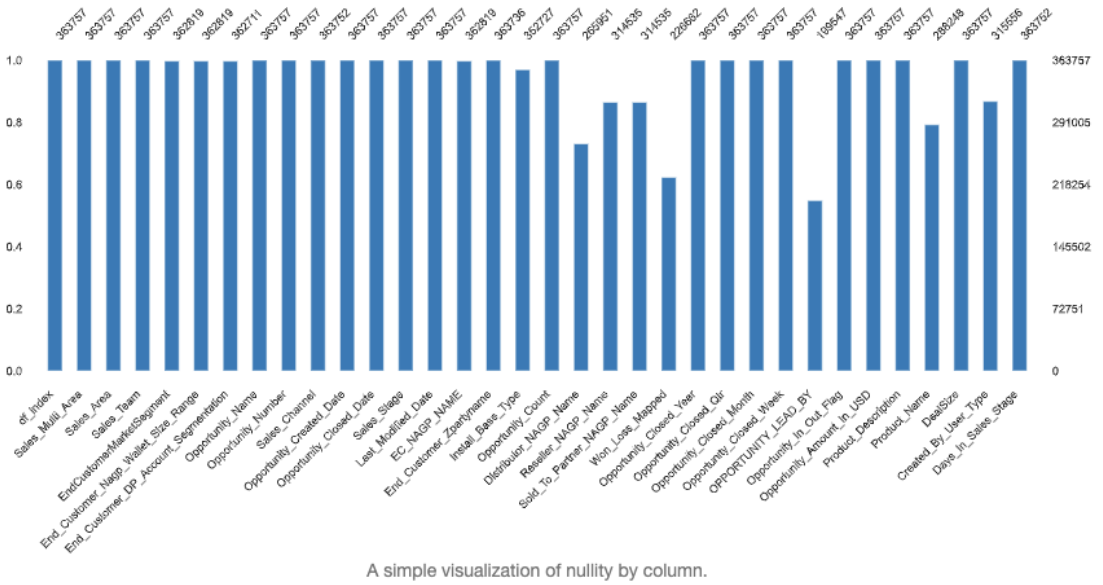

Figure 4.1: Missing Data - Split by Features

- With human based input, We can expect mistakes in Product names. So as part of this step, mapping is done to ensure correct product names and product classifications are reflected all across the dataset.

- Since the pipeline is used for also testing purposes, It contains some test opportunities. These are marked out as test under description and they are filtered out from the dataset as well.

4.3.2 Adding New Features

As part of feature engineering, we create new features from already existing features with the assumption that they will help produce better output. Features created reflect business knowledge and the cyclical aspects of an opportunity. Table 4.3 reflects the new features added and their description. The target feature "Label" is also created in this step which conveys whether a deal was won or lost, which is the expected outcome.

As a final step we keep the newly created features and drop unwanted features that convey the same information, i.e., duplicate features.

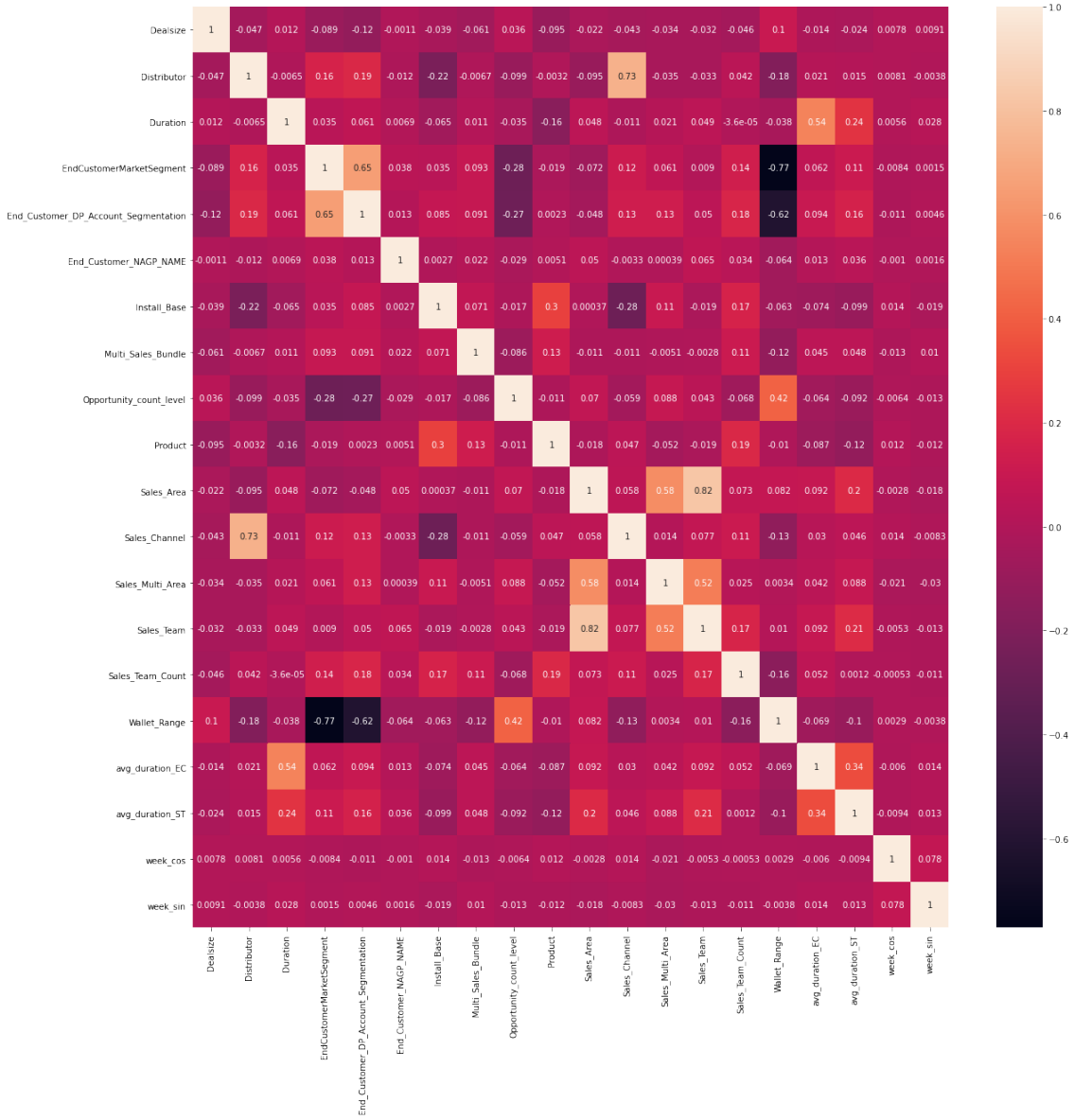

Figure 4.2: Heatmap - Displaying correlation

Feature	Description
Duration	Time taken to create and close an opportunity (in days).
Install Base	Classification between New Customer,Technology Refresh or renewal based customers.
Distributor/Reseller/Sold to Partner	Separate features explaining the medium of the sale
Quarter/Month/Week	Date features derived from the closing date of an opportunity.

Table 4.3: New Features Description - Details

4.3.3 Categorical Encoding

Categorical Encoding is highly important to improve the model's performance, and also the ability to translate the same level of knowledge from categories to encoded values is important. This essentially means that the model needs to be able to interpret the meaning of the categorical features correctly, which leads us to effectively use the categorical data we have at hand. When it comes to categorical data, there is no one possible pathway to success. It is a trial and error method and one needs to experiment with different encoding techniques to see which improves the performance of the specific machine learning model to use. There are some commonly used Categorical Encoding techniques such as One Hot encoding, Nominal encoding and Ordinal encoding.(13) In particular, One hot encoding is widely used when the categorical features doesn't convey any information about order and doesn't need to be represented in an orderly manner. As mentioned in section 4.2.1, the training dataset consists of High cardinality features as well. The problem with High cardinality features is that they come with some difficulties owing to the different values present in them. It's difficult to apply the common techniques like one hot encoding or label encoding as they lead to introduction of sparsity in the dataset or lead to high number of unique values which fails to convey the meaning of these values. (14) To combat encoding high cardinal features, there are some advanced encoding methodologies such as Weight of Evidence encoding, Target Encoding, Mean encoding etc. Also, we have some Machine learning algorithms which can automatically encode the categorical values as per their own encoding scheme, such as the XGBoost encoder and CatBoost encoders.

With all this in mind, we apply different encoding schemes to the categorical variables and compare their performance on different classification algorithms. Upon doing this, the results show that applying CatBoost encoder provided better performance in terms of accuracy and f1 score for our dataset. Table 4.4 shows the performance metrics after the application of the CatBoost encoding technique on different classification algorithms.

ML Algorithm	F1	Accuracy
RandomForest Classifier	0.77	0.77
XGB Classifier	0.76	0.77
CatBoost Classifier	0.77	0.77
DecisionTree Classifier	0.71	0.72

Table 4.4: CatBoost Encoding Performance

4.3.4 CatBoost Encoding

CatBoost encoding is a target based categorical encoder. It is similar to the target encoding, but avoids the problem of target leakage by means of an ordering principle. CatBoost in general computes target statistics based on the ordering principle. The method takes a random permutation of the training samples and for each sample it computes the target statistics by using all the available history. This essentially means the target statistic for a categorical feature is calculated only from the observations before it, which helps to overcome the problem of target leakage. (15) (16)

CatBoost encoding takes into account prior value, which is common practice and helps to reduce noise from low frequency categories. For binary classification methods, the prior is a priori probability of encountering a positive class. This calculation methodology makes it to be flexible when encountering new features as well.(17)

$$\frac{(TargetSum + prior)}{(FeatureCount + 1)}$$

where TargetSum reflect the sum of the target for that particular feature value, prior is a constant calculated at the start and FeatureCount is the total of categorical features observed with the same value as the current one.To understand the encoding methodology even clearer, we introduce an example. Suppose we have a categorical feature "car" with values like ['BMW','AUDI','AUDI','BENZ'] and target column as "Bought" with values ['1','0','1','1']. From this as an initial step, we calculate the prior to be 3/4 = 0.75. For the category value 'AUDI', the target count will be 1+0 = 1 and the feature count will be 2 as there are two instances of that value. So the encoded value for 'AUDI' will be (1+0.75)/(2+1) = 0.58. This is how the encoding is calculated in general if we consider this as one permutation.

The CatBoost encoding is implemented using the CatBoost library along with the CatBoost classifier. There are several other implementations, with the most common one being using the category encoders library of python, which provides implementations for a wide range of encoders.

After carrying out all the processes as mentioned in the previous sections, The data is now ready to be used for building the model. ~~The~~ table 4.5 shows the final list of features that were present before applying feature selection techniques on them.

Group	Features
Date Based	Quarter,Month,Week,Duration
Product Based	low/high flash, low/high PCS, low/high Hybrid Cloud, PS , Renewals, Others
Sales Territory	Sales Area, Sales Multi Area
Customer	Customer market segment, Wallet Range, Install Base
Opportunity Detail	Distributor,Deal size,Sales channel,Created by user type,Sales Team, Amount

Table 4.5: Final list of Features

4.4 Feature Selection Approaches

Feature Selection is an important step in machine learning. It helps to identify important features and also helps to disregard unimportant features. This contributes to improvement in model's performance. The literature on feature selection methodologies states three different types of feature selection approaches(18)(19). The three common approaches are

- Filter Methods: The filter methods are based on statistics and use an approach which is completely independent from using a machine learning algorithm. The filter method uses rank ordering method for variable selection. Techniques such as Chi-square test,Pearson correlation criteria , Mutual information etc. are part of the filter methods. Since they involve statistics and are based on ranking methodology, the filter methods tend to be computationally faster even with high dimensional datasets due to their tendency to scale and also avoids the problem of over fitting. The main disadvantage is that it considers each feature individually and scores them which can lead to degradation of the classification performance as it doesn't take into account the interaction of variables.

- Wrapper Methods: The wrapper methods are another type of feature selection techniques. The wrapper methods are dependent on the machine learning algorithm to select the features. They select the features that help to improve the classifier's performance in terms of classification accuracy and this makes them classifier dependent methods. Recursive Feature Elimination (RFE), Backward Feature Elimination (BFE) and Boruta (20) are some of the most common wrapper methods. The wrapper methods are computationally expensive when compared to filter methods. They become more expensive when the size of the dataset increases. Although the wrapper

method takes into consideration the interaction between variables, it can still cause over fitting as it strives to improve the classifier's performance.

- Embedded Methods: Embedded methods are a combination of wrapper and filter methods. They combine the qualities of both and are implemented by all machine learning algorithms as an in build feature selection approach. RandomForest, CatBoost, Decision Tree, Linear SVM are some of the algorithms that have their own inbuilt feature selection capabilities. The benefits of the embedded methods are the models results are interpretable, and they generalize better.

As mentioned, all of these different methods have different advantages and disadvantages. With the help of the correlation heatmap shown in figure 4.2 and also as mentioned in subsection 4.3.1, we remove certain features that were highly correlated and also those that were conveying the same information. As part of the thesis, Wrapper methods like Recursive Feature Elimination and Boruta Algorithms were tried out, but they were not chosen due to their time complexity. With a larger training data, we experienced the feature selection task using wrapper methods to be time-consuming and computationally expensive.

For the thesis, The approach was to use a combination of Filter and Embedded methods and choose the feature with the highest votes, i.e, gets picked by the different algorithms as important. The methods used are listed as follows

- Chi Square Test - It is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

- Random Forest - RandomForest classifier is used for this method. Random Forest uses its inbuilt feature selection to evaluate the model, and it works on the principle of how pure each of the buckets containing similar data but different from one another. For classification problems, the model uses Gini impurity or Information gain/entropy.

- CatBoost Classifier - CatBoost Classifier also is an embedded method with its own inbuilt feature selection. It works on getting the feature importance by means of understanding the weight of the leaves in the model.

Table 4.6 shows the feature scores along with the different feature selection algorithms. "1" conveys it's an important feature, and "0" means it can be removed. The "final

score" column shows us the overall scores for each feature. The idea initially was to take the features which scored greater than or equal to two and apply them to the classification algorithm. Results from that approach showed that removing all features with one as the score resulted in a volatile model performance. So after keeping and different features, we decided to remove the product grouping features as those were the ones producing the changes in performance. The features is_PS, is_Renewals, is_High_Flash, is_low_flash,is_high_PCS, is_low_PCS and is_others were removed. Features such as Deal size, Install_Base, Distributor, Created_by_user_type and End customer market segment were kept even though they had a score of one.

Features	RF	Catboost	Chi	Score
Duration	1	1	1	3
Wallet_range	1	1	1	3
Sales_Team	1	1	0	3
Sales_Area	1	1	0	3
Month	1	1	0	3
is_PS	0	0	1	1
is_others	0	0	1	1
is_Renewals	0	0	1	1
Sales_channel	1	1	0	2
Sales_multi_area	1	1	0	2
Sales_channel	1	1	0	2
Deal size	0	0	1	1
install_base	0	0	1	1
is_high_flash	0	0	0	0
is_hybrid_cloud	0	0	0	0
is_low_flash	0	0	0	0
is_low_pcs	0	0	0	0
endcustomermarketsegment	0	0	1	1
distributor	0	0	1	1
created_by_user_type	0	0	1	1
is_high_pcs	0	0	0	0

Table 4.6: Feature selection Approaches

5

The Hybrid Approach - Machine Learning with Simple Maths

This chapter explains the approach taken to solve the problem at hand, the design of the model where we try to distinguish between how the training and the active pipeline works and further details about the workflow of the model.

5.1 Overview

There is always an end goal which is to be reached with every research. For this work, The end goal is to predict whether the opportunities will be won or lost and to formulate an expected forecast value at the Sales Multi Area level. With Sales Multi Area level, the expectation is to get the forecast values at an EMEA, Americas and Asia Pacific level. Additionally, we provide visualizations to explore the historical data and the actual progress of the Sales. To facilitate this, the design consists of three components.

- Machine Learning : At the heart of the design is the Classification Machine learning algorithm - Predictor.There are different ways to approach this prediction problem. It can be considered as a Classification or as a Regression problem. As mentioned in Chapter 3, There have been related work done considering the problem to be either of the above two. With that said, We take the Classification approach instead of the regression approach. We treat the target as labels, with values as Won or Lost. This provides the answer to the first part of our goal, which is to predict whether an opportunity will be won or lost.
- Mathematical Approach : The next stage is to formulate an accurate forecast value at the Sales Multi Area level. For this, we consider only the won deals from the output of

the machine learning model. The problem is that the sale value for some opportunities are not accurate. There are chances for opportunities to have zero or very high values in the sales pipeline. Although we exclude opportunities with abnormally high values in the preprocessing stage, we cannot do the same for opportunities with zero values. So when we aggregate the sales value at a Sales Multi Area level, it leads to higher or lower forecast value, which in turn triggers the major problem of Over forecasting or Under forecasting. This will also be explained in detail in the later sections of this chapter. To overcome this, we employ a simple mathematical approach which takes in as input the output data from the machine learning model and the Historical Sales data from the past, which only contains deals that got converted into revenue for the company. With the help of a weighted average, we calculate the forecast values for each multi area, with which we find a solution for our second part of the end goal.

- Visualizations: The final component is the Visualization. The historical sales pipeline needs to be monitored to check its hygiene and health aspects. Also, we need a medium to communicate the output of the machine learning model. The solution to the third part of the end goal lies with Tableau. We use tableau to create dashboards that convey the information required to monitor and assess the sales pipeline.

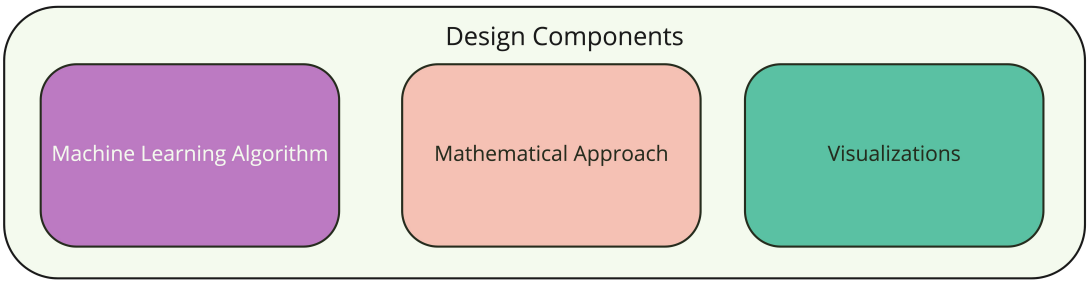

Figure 5.1: Design Components

5.2 Machine Learning Algorithm - Classification Model

Our approach falls in the line of Binary Classification, where we classify the opportunities into either won or lost stage. Binary Classification is the task of classifying elements into one of the two available groups (21).Various Binary Classification algorithms have come into picture ever since the increase in the usage of machine learning to solve real life problems.

Some ideal examples of binary classification problems are marking an email as spam or not, identifying whether a person has a disease or not etc. Different types of classifiers are in practice these days such as linear based classifier, Tree based classifiers, Support vector machines and much more complex classifiers based on neural networks. Ensemble methods (22) and Gradient boosting algorithms are also a welcome addition to the list of classifiers. Some of these classifiers also extend support to Multiclass classification, where the elements are classified into more than two groups. An ideal example for the multiclass classification would be the prediction of weather type such as sunny, windy or rainy. The underlying concept behind binary classification (21) is that the elements are applied to a classifier and the classifier calculates a prediction score which potentially reflects the ability of the element to be classified as the positive class. The algorithm after predicting the score compares it to a classification threshold and if the predicted score is greater than the classification threshold then the element belongs to the positive class and vice versa. For our thesis, we take a closer look at Gradient boosting algorithms, as they were the ones which performed superior in terms of classifying opportunities as won or lost.

Gradient boosting algorithms train a sequence of weak models, where each of them compensate for their predecessor's weakness.(23) The idea behind gradient boosting is to choose a weak model which can be a decision tree or a regression model and build on top of it. Most popular types of Gradient boosting algorithms are AdaBoost, XGBoost, LightGBM and CatBoost. XGBoost, LightGBM and CatBoost are the most widely used boosting algorithms as they give a greater model performance and are faster compared to the other algorithms present. For the prediction purposes, we used the CatBoost Classifier - a supervised classification algorithm based on Gradient boosting. The motivation behind the idea to use CatBoost instead of other algorithms was because of the following reasons.

- Firstly, The model gave a superior classification accuracy, and it produced higher scores in terms of f1-score when compared to the other models as shown in table 4.4.
- CatBoost also offers support to categorical features. Since our data is highly categorical, this ensured good results.As discussed in the Challenges section 7.1,There were also some problems faced with respect to categorical data that were handled well by CatBoost.

5.2.1 CatBoost

CatBoost is a machine learning algorithm that uses gradient boosting on decision trees. It was developed as an open source machine learning algorithm library by Yandex in 2017.

The CatBoost Classifier is one of the implementations of the CatBoost algorithm, which aims to solve any Binary Classification and Multiclass Classification problems. The CatBoost algorithm is stated to be performing highly with respect to other gradient boosting algorithms like XGBoost, LightGBM etc. There have been benchmark tests performed with CatBoost and other gradient boosting algorithms which have conveyed the same. This makes us to understand what makes CatBoost better when compared to other algorithms.

Literature based on CatBoost from the authors of the algorithm explains in detail how CatBoost works (16)(15). CatBoost as the name states is the acronym for Categorical Boosting and the algorithm specifically performs better in terms of handling and encoding categorical features with different methodologies when compared to others. The algorithm makes use of permutation techniques, One hot max size and target based statistics to handle the categorical features. The CatBoost algorithm divides the dataset into different subsets and encodes the categorical values using one hot encoding, target statistic or count based depending upon the specifications of the model. The algorithm avoids overfitting and target leakage by the concept of ordered boosting, for which a hint of what it does is provided in the section 4.3.4. The other advantages of CatBoost is the ability to produce great quality without parameter tuning, along with faster training and prediction time. This was also critical for us as we worked with limited resources for the thesis. The thesis implements a default CatBoost classifier with the learning rate specified at 0.1 with all other features set to their default values. This also highlights the aspect of the algorithm where it can produce high quality without parameter tuning, as it performed better than all other classifiers on the B2B sales data.

5.3 Mathematical Approach

The mathematical approach is taken to get an accurate forecast value at the Sales Multi Area level, which is one of the goals of the thesis. A key motivation to add this additional layer on top of the machine learning model was discussed in brief in the section 5.1. As mentioned in section 5.1,the sales value associated with an opportunity determines the accuracy of the forecast that can be formulated. Listed below are the key factors to pick up this approach:

- The sales value can be zero in the sales pipeline data, but can have a significant value in the actual sales data. For example, an opportunity "XYZ" of a customer "ABC" can have 0 as its booking amount in the sales pipeline. This opportunity is a won

deal. So there is an entry which is made in the actual sales dataset which houses only deals that are won by the company and this is absolute truth in terms of calculating the sales at the company. The same opportunity "XYZ" has a value of 200,000 euros in the actual sales dataset. The predictor uses the sales pipeline data and even if the predictor identifies this deal as a won deal, the forecast formulated at the end cannot account for the difference of 200,000 euros which leads to an under forecast. Similarly, the opposite of this use case is also possible that can lead to over forecast.

- Usually, the classification accuracy of a model can be improved with the help of a confusion matrix. By varying the threshold level, we can reduce or increase the True positive or True negative. This is usually the case when we try to apply machine learning in areas such as spam filtering or cancer detection, where we need to be account for the maximum of true positive cases and there can be occasional misses. But the same concept cannot be applied in the B2B sales forecasting area, as this again creates the problem of over/under forecasting. If we account for either one, it leads to a biased output.

To overcome these, we employ the mathematical approach. The input to this is the output data of the machine learning model as stated in section 5.2, the actual sales data of the company over the years post 2020. Figure 5.2 shows the steps on how the mathematical approach is performed. The approach places weight on the historical actual sales data. It considers the previous year and gives it a higher weight to account for recent trend, and it also considers the value from two years back and places a weight on it to account for consistency. Figure 5.3 explains the steps with an example.

5.4 Visualizations

The visualizations constitute the final component in the design of this model. Tableau, the popular business intelligence tool, is used to create the visualizations. The visualizations are aimed at providing insights into the sales pipeline's Hygiene and Health. Sales pipeline keeps constantly changing with respect to time. There is always opportunities flowing into and out of the pipeline. Since this is used to formulate the forecasts upon which the companies build on their plans for the upcoming periods, it's essential to monitor the sales health and hygiene aspects closely.

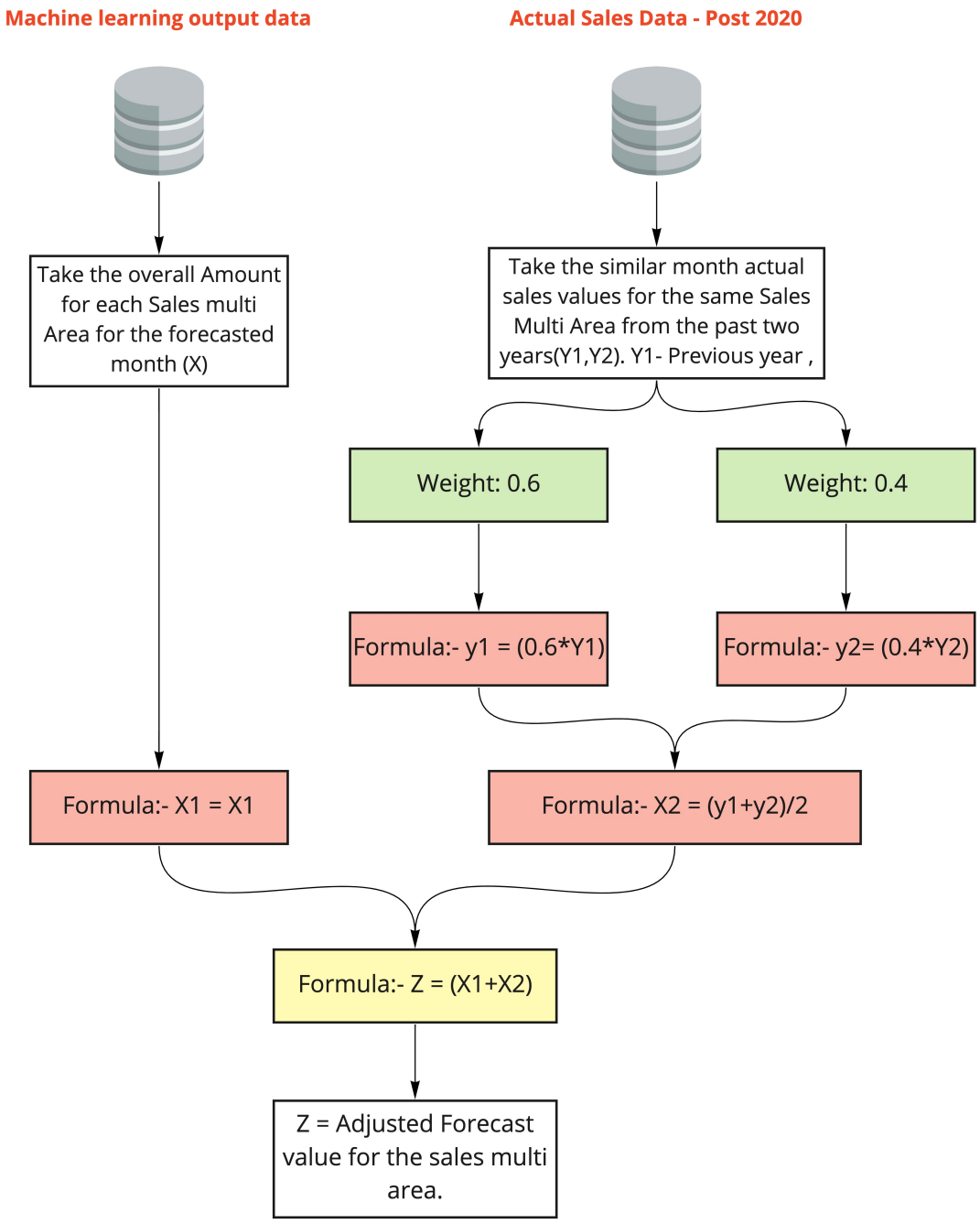

Figure 5.2: Mathematical Approach to formulate forecast - Steps

Hygiene of the Sales pipeline refers to the process of keeping the data as accurate and up to date as possible. This refers to the process of checking the opportunities for the following:

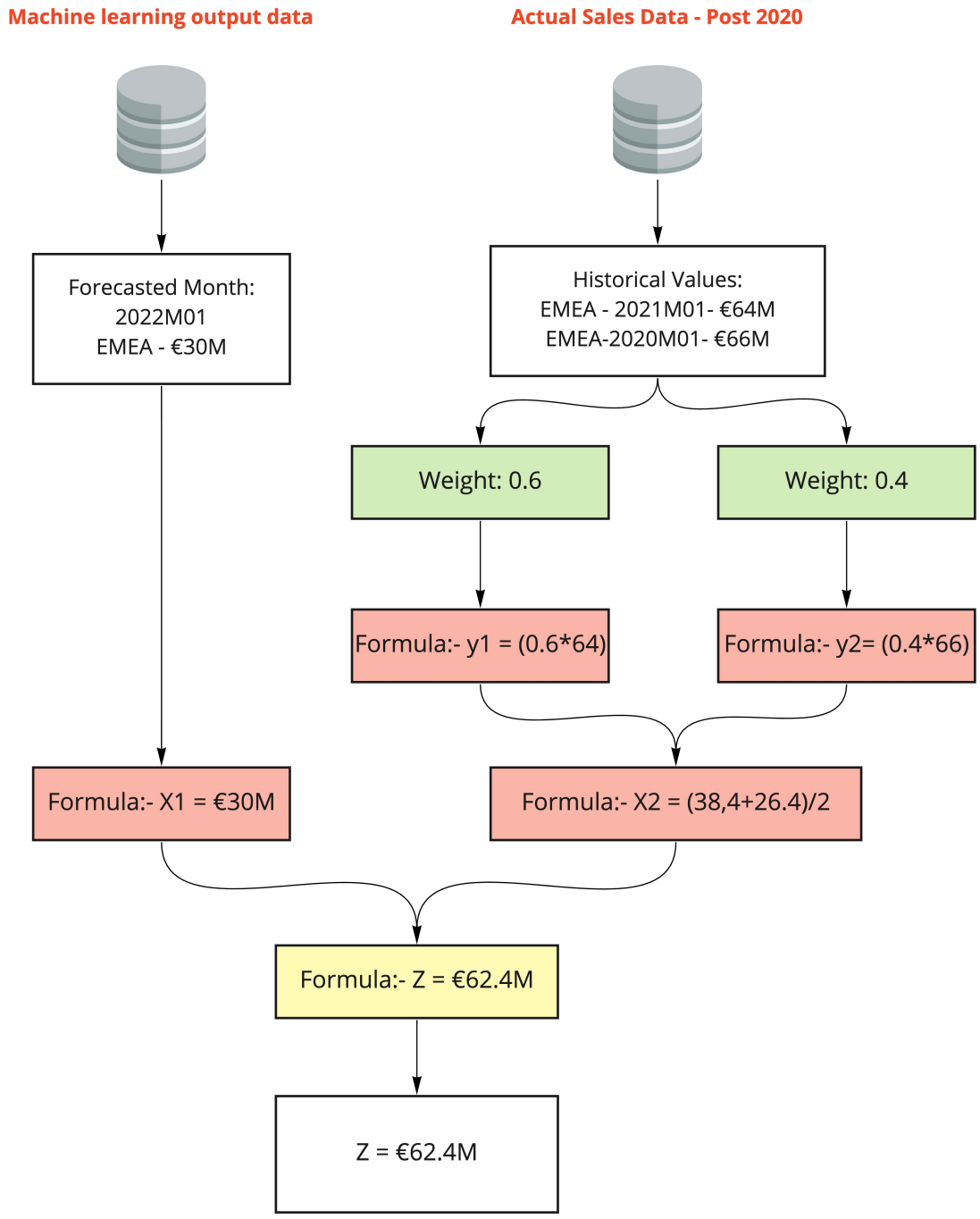

Figure 5.3: Mathematical Approach to formulate forecast - Example

- Should reflect correct close date.
- Should be tagged with the correct products present in the product portfolio.

- Should reflect reasonable sales value for opportunities.
- Ensuring the opportunity details get updated at each stage or upon any change in status of the opportunity.

Maintaining a good hygiene will result in improved forecast accuracy, helps lay the foundation for analytics, and also helps sales representatives learn about the opportunities. Although it has advantages, Hygiene is one of the hardest things to track, as there is a common misconception that spending a lot of time on entering and updating the records leads to less time for the sales professionals to carry out sales activities. Also, there is an impression that it doesn't help the sales teams, which leads to failure in maintaining good hygiene levels.

Similarly, Health of the Sales pipeline refers to the assessment of the pipeline to identify warning signs or achievement's in one's business. The health of the sales pipeline is monitored to ensure the following:

- To ensure that there are enough opportunities in the pipeline to reach the estimated value in the annual operating plan.
- To understand the success rate at particular points of time - Win/Loss monitoring of opportunities,
- Duration taken to close opportunities on an average.
- Identifying opportunities that slipped, i.e, those deals that were to be completed in a promised time period but didn't materialize and instead they were shifted to a newer time period.
- Understanding the reasons behind why an opportunity is lost.

Hygiene and Health are related. If there is a good hygiene for the sales pipeline, then it will enable to drive good health analytics on a daily/weekly/monthly basis which can help drive discussions and help improve the organization's business capabilities.Along with this, we also need a visualization to communicate the output of the machine learning model and the output of the mathematical approach. Tableau is used to develop the dashboards, which will help provide insights into the three different tasks at hand. We design four different dashboards that help to solve the above problems.

5.4.0.1 Predictor Output Dashboard

This dashboard is composed of two views. The first view shows the output of the mathematical approach. As mentioned in section 5.3, the mathematical approach helps to formulate the forecast value. So this view displays the forecast value aggregated and adjusted at a sales multi area level. The second view lists the opportunities and provides the predicted output associated to each of them. Each opportunity is displayed with its ID, Name, Multi Area, Area, Customer name, Product name, booking amount and its predicted label. Both of them are plotted in the form of tables. It only reflects the details for the selected forecast month.

5.4.0.2 Pipeline Hygiene - Active

This dashboard is used to display the details regarding the hygiene of the active pipeline. It consists of four views. The first view is a circle view that identifies Opportunities above a certain threshold sales value, which helps in filtering the opportunities with unreasonable sales values. The second view is a bar chart with each bar representing the count of opportunities with zero as their booking amount and are grouped by the products present in the product portfolio. The third view is again a bar chart which shows the amount of missing values present in critical columns which are key to monitor the sales pipeline. The final view is a tree map which shows opportunities with a close date lesser than their creation date.

5.4.0.3 Pipeline Health - Active / Past

Two dashboards are used to represent the pipeline health for the Active sales pipeline as well as the past sales pipeline.

- Active Sales Pipeline Dashboard: For the current month, This dashboard explains the deals in the pipeline, which gives a sense of understanding about the business for the month. The dashboard churns out information such as the count of opportunities won and lost in the period which are looked at from different dimensions, lists the opportunities by different stages of the opportunity lifecycle as shown in figure 2.1 and finally shows the sales for that month over the years to give an idea of how it's doing year over year.
- Past sales Pipeline Dashboard: With emphasis to reflect on the performance over the past years, this dashboard helps to understand how functional the sales teams were

and how the sales took place. So this primarily shows the performance of Sales teams using the number of opportunities they closed, lost and their conversion percentages. This also helps to take a look into what path the sale took, whether it was directly sold or whether the sale happened via different resellers and distributors. The final part of the dashboard shows the pattern of won/close deals over the years for each month, with which one can get a clear understanding of what kind of sales pattern to expect.

Thus, we find the solutions for the three part goal which was explained in the section 5.1. In the following section, we explain the two different pipelines used to achieve the experimental portion of the explanations in the previous sections.

5.5 Model Flow

The architecture of the model as shown in figure 5.4 takes the approach of having two different flows - the Training flow and the Active flow. The following subsections talk in detail about the different pipelines and how we set them up.

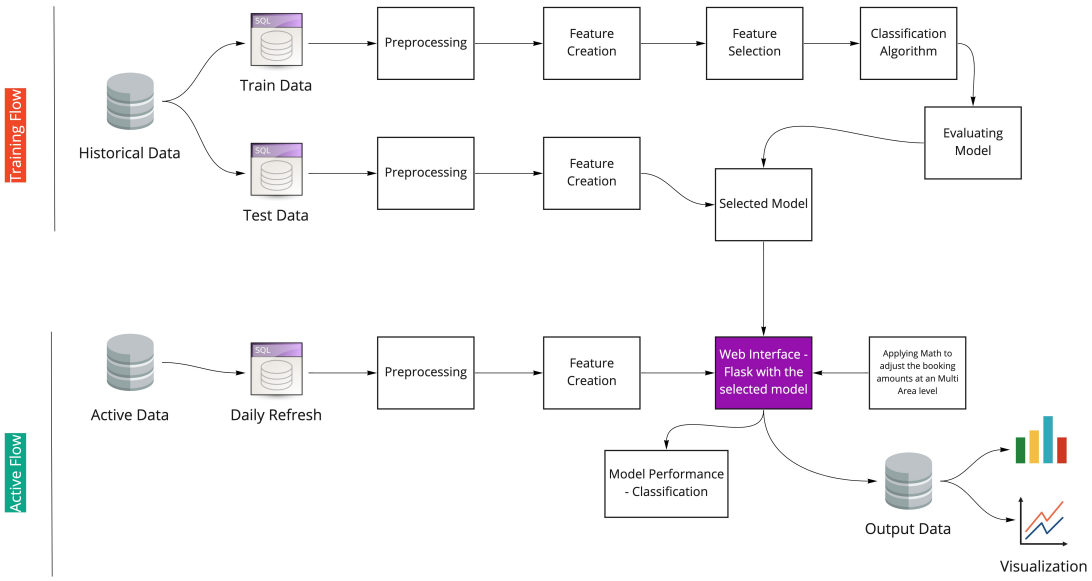

Figure 5.4: Predictor Architecture

5.5.1 Training Flow

The training flow is made up of all the different components mentioned in Chapter 4. The historical data is queried from the database, which is then split into train and test data for applying machine learning techniques. The training flow consists of two branches as shown in figure 5.4 depending on the data being used. Branch one uses the train data and performs the preprocessing as mentioned in section 4.3.1 and feature creation as mentioned in section 4.3.2. Post this feature selection is applied as mentioned in section 4.4 which helps us to choose the best features to use for the predictions, and we also stated that we take some features which are not considered important by the feature selection techniques. The filtered data is then fed to the different classifiers from which CatBoost was selected. Similarly, the feature creation and preprocessing steps are carried out on the test data and this is directly passed to the CatBoost classifier. This pipeline is created, but there is no scheduled run to retrain the models. Evaluation of metrics is taking place constantly, and the process of retraining is triggered manually only when there is a drop in the performance or if there are some changes on the data side. The implementation is completely done using Python and the data querying was done using SQL. The libraries of Python that were majorly used to bring this flow to working mode are listed in the table 5.1.

Function	Libraries
Data Loading	pyodbc
Data Preprocessing	Pandas, numpy
Feature Selection	sklearn
Classification Algorithm	sklearn, catboost, xgboost, lightgbm
Monitoring	Evidently.ai

Table 5.1: Python Libraries

5.5.2 Active Flow

The production based model which is used by the team makes use of the active flow. The active flow uses the new data flowing into the system. All the components in the active flow are packaged into a flask application which is run on a production server. An interface is present in the front and when the user clicks on "Generate forecast", it triggers the

rest of the processes in the background. The data flowing into the system is stored in a table in the database and the table is truncated and created every month to reflect the Sales pipeline data for the current month which is to be forecasted. A scheduled SQL job keeps this table updated with the new data. This data is then preprocessed, and the new features are created as well, and this data is fed to the CatBoost classifier selected from the training flow. This predicts the outcome of the opportunities and this in turn triggers the mathematical approach which calculates the forecast value at a Sales multi area level. Finally, the output data of the predictor as well as the mathematical approach is written back to the database. This output table is used in tandem with the historical sales pipeline, historical sales data and the current month's pipeline data to create the dashboards in Tableau. The classification model is periodically evaluated every month by an automatic performance profiler to identify any deterioration in the model performance. Also, this helps to make a decision on whether the model needs to be retrained or not.

This whole process takes around 2-3 minutes for each click on the "Generate forecast" button as shown in figure 5.5. The whole flow is implemented using python and the whole web application is containerized using docker. The web interface is built using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS). The container created is deployed into production using Kubernetes and is monitored constantly to support in the event of a crash.

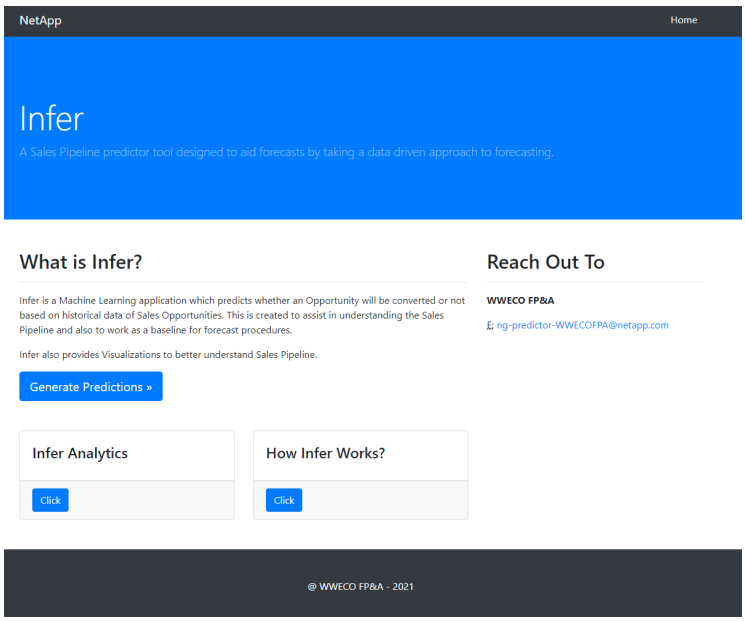

Figure 5.5: Infer - interface

6

Experiment & Evaluation of Results

The experiment carried out is discussed in detail in this chapter. Evaluation of Results is a key component that helps us to understand the outcome of the experiment and the trust level behind the results. This chapter talks about the classification metrics we looked at and the accuracy behind the forecast results. Also, it takes a look into how effective the dashboards were in terms of discussions among team members and the action points they could derive from them.

6.1 Experiment

From the training flow, The explanation of the training data is already provided in section 4.2. We perform a K-Fold cross validation (24) with K as 10 to ensure whether the model is consistent in terms of performance and also to avoid the problem of building a model that is overfitting/under fitting. There is a high volume of data and this drove the motivation to use a K-Fold cross validation to evaluate the model performance. The commonly used approach which is the ratio based train test split method can be used here, but it always has the problem of what kind of data it takes for the test split. Although, the split can be controlled in terms of balancing the data samples taken by the method, yet we can't really say that the model is generalizing to all samples of data. Maybe there is chance that the model performs well for the divided data or maybe not and this makes us to use K-Fold cross validation which helps minimize this uncertainty. Post this split, the execution of the training flow happens, where we prepare the data and fit different classification algorithms and evaluate the performance of the different classifiers.From this we select the best classifier which will be used as part of the active flow. The entire process is implemented as a pipeline using the sklearn library. However, we cannot make use of the

same functionality for CatBoost due to support issues. So we handle the cross validation using CatBoost individually with the model's inbuilt capabilities.(25) The training flow also consists of a test data. With K- fold cross validation, we get to know the overall model performance on the training dataset and this only gives us an idea of how the classifier performs on the dataset in terms of predicting the opportunities as won or lost. Although it ticks the evaluation criteria for one of the thesis goals, we also need to understand the forecasting accuracy, which is the main expectation. So, as shown in table 4.2, we have a separate dataset which contains data only for the fourth quarter of financial year 2021.We make use of this dataset as it makes way to compare the obtained forecast values and the predictions with the actual sales values for this period. At the time of doing the thesis, the fourth quarter has already passed, So we took all closed and won deals from this quarter and used them as the test set. Although the test set is formed differently compared to usual approaches where we just use the Train Test split methodology, this way of formulation was required to obtain a clear picture of how the process was performing.

6.2 Evaluating CatBoost Model Performance - Chosen Performance Metrics

As mentioned in section 5.2, Different classification algorithms were applied to the get the right classifier for our data. Since we employed the K-fold cross validation method, we average the different estimates of the classification metrics for each classifier and use that as the scoring metric to finally select the best model. Table 4.4 already states the scores of the different classifiers, and the section 5.2 discusses why we chose CatBoost as the best fit for our use case. This section evaluates the performance on the test data from the training flow, which is described in the section 6.1.We formulate the performance metrics based on the three different components shown in figure 5.1 to evaluate the model performance.

6.2.1 Classification Metrics

Evidently AI - a python library (26), helps to evaluate the model performance on both the active flow and the training flows (section 5.5). The library provides different automatic and inbuilt reporting techniques to help monitor the machine learning models in production as well as in development. This can be only used on tabular data and not any other forms of data.We also make use of the CatBoost's inbuilt evaluation tools to obtain some additional information about the performance of the classification model. The evaluation happens automatically as part of the active flow, but it is trig erred manually in terms of the

training flow. We make use of the classification report generated by Evidently and take the data points from them to evaluate the model. The report is based on the performance of the model on the test dataset of the training pipeline shown in figure 5.4. The class representation of the test dataset is shown in figure 6.1 which shows a fairly balanced dataset. The amount of lost opportunities is a bit higher than the won opportunities, which aligns perfectly with the business understanding. From that report, we look at the

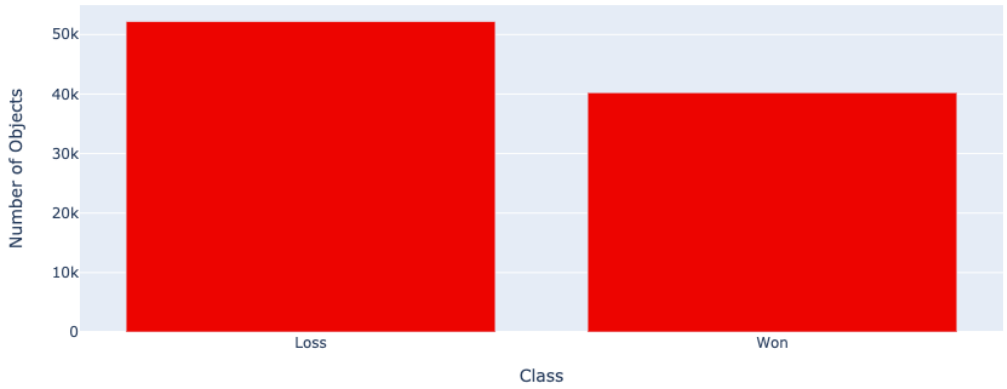

Figure 6.1: Class representation of Test Data - Training Flow

following metrics listed below. As mentioned in the previous sections of the paper, the aim of this classifier is to predict both the won opportunities and lost opportunities efficiently. There cannot be a scenario which can be accepted in which the model performs extremely well for True positives and doesn't perform well in identifying the true negatives. This is because the ultimate aim is to predict the number of opportunities which will be won or lost, and business needs to have a proper understanding of both in order to work towards improving in the future. Accounting for either True positives or True negatives will lead to the problem of under/over forecasting and if that happens, it will lead to the downfall of the entire idea behind the project. Keeping that in mind, we identified the following metrics to evaluate the classifier performance.

- Confusion Matrix: Confusion matrices (27) are the most commonly used evaluation metrics when it comes to determining a classifier's performance. It gives a summary of the predicted results of a classification by giving the listing out the number of correct and incorrect classifications the model made. The CatBoost classifier's confusion

matrix on the test dataset of the training pipeline is shown in the figure 6.2. From the confusion matrix, it's quite clear that the model is making a good generalization when it comes to predicting the correctly won(29091) and lost deals(42844). Also, from the confusion matrix, we can observe the opportunities that were misclassified. The conclusion is that the model minimizes the impact of false positives when compared to the impact of false negatives, but still there is a significant amount of opportunities getting misclassified.

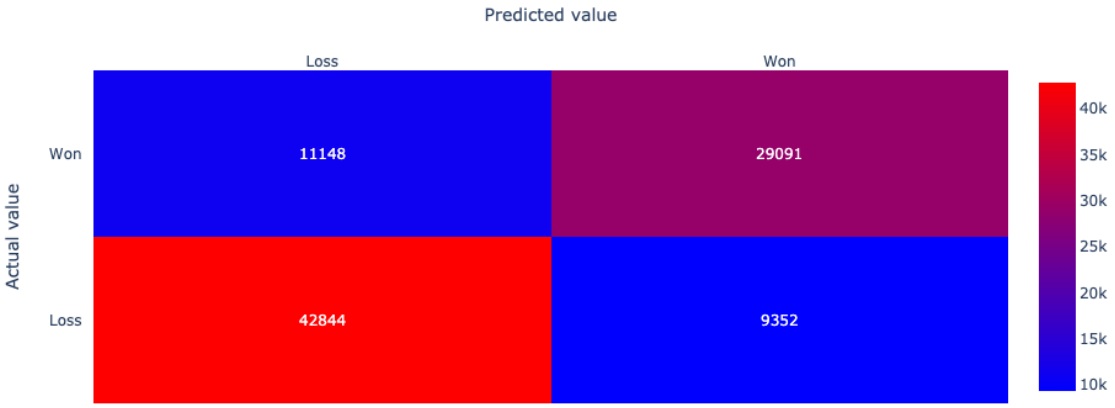

Figure 6.2: Confusion Matrix of Test Data - Training Flow

- Precision, Recall, F1-Score and Classification Accuracy: The classification accuracy measures the ratio of correct predictions over the total number of instances. The precision and recall measures are also taken into consideration, where precision gives the ratio between the true positives and all the positives present in the data and recall is the measure of us identifying correctly the true positives.F1-Score is the harmonic mean of the precision and recall. (28) The equations 6.1, 6.2, 6.3, 6.4 represent the way in which the terminologies are calculated.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{6.1}$$

$$Precision(P) = (TP)/(TP + FP) \tag{6.2}$$

$$Recall(R) = (TP)/(TP + FN) \tag{6.3}$$

$$F1 - Score(F1) = 2 * ((P * R)/(P + R)) \tag{6.4}$$

The values recorded for the classifier model are shown in table 6.1. To account for both the precision and recall, we consider the F1 score, since our goal is to have a good precision as well as recall. The model performance is overall at the 77% mark for all the performance metrics discussed, which is acceptable in terms of classifying the win/lost opportunities. Figure 6.3 shows the quality metrics by Class for the test dataset - Metrics by label value.

Figure 6.3: Quality Metrics by Class for Test Data - Training Flow

Metric	Score
Accuracy	0.778
Precision	0.775
Recall	0.772
F1 score	0.773
AUC Score	0.772

Table 6.1: CatBoost - Classification Performance Metrics

- AUC-ROC: AUC-ROC Curve (27) is a performance metric which gives an overall ability of the classifier to classify as 0's as 0 and 1's as 1. Unlike the other metrics mentioned above, which are more inclined to one of the classes, AUC gives a clear picture about the ability of the classifier in prediction. The higher the AUC, the higher is the model's capability to distinguish between the classes.From table 6.1, the AUC score computed is 0.772. Computing the ROC curve for the CatBoost classifier is an inbuilt capacity of the CatBoost library of python, and it helps to compute the AUC curve over multiple iterations to obtain the best fit. We use the loss function as "AUC" to get the ROC curve from the classifier as shown in figure 6.4.

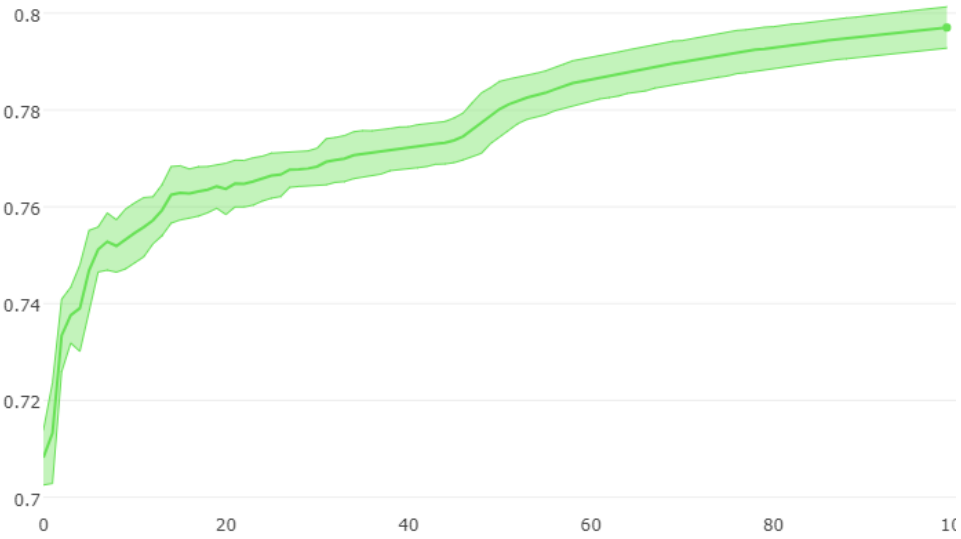

Figure 6.4: ROC Curve of the Test Data - Training Flow

- Model Explainability: The model's explainability is shown using SHAP values (12). SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value. Figure 6.5 shows the features and their SHAP values which lead to influencing the predictions of the Classifier.

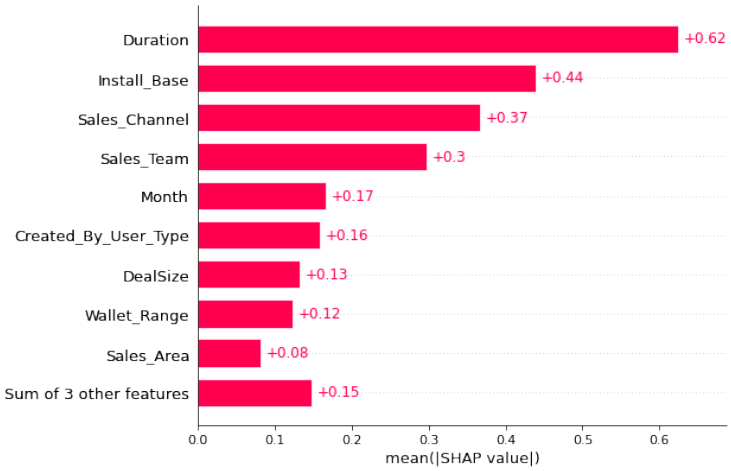

Figure 6.5: CatBoost Explainability of the Test data - Training Flow

6.2.2 Forecast Performance Metrics & Visualizations

Post the classification and with the metrics in the above section, we find the best classifier for our problem. As shown in the figure 5.1, our process of estimating the forecast consists of a mathematical approach on top of the classifier output, which finishes off with visualizations to support the process. To assess both the performance of the final forecast and the visualizations, we formulate two metrics:

- Forecast Value accuracy: The forecast value accuracy is calculated as shown in equation 6.7. The value C is the adjusted forecast value calculated via the mathematical approach mentioned in section 5.3. We also know the actuals sales value for the data present in the test set which is taken as the value for D. We calculated the accuracy of our estimated forecast to be 83% which means our C value only missing from the D value by a margin of 17%. Thus, our results show that the predicted and formulated forecast can be counted as reliable. Although there is a 17% margin, this is estimated in comparison with the actual sales in that time period, which doesn't always need to tally with the forecast and is dependent on the other factors such as the business of the company and has other external factors as well. Nine out of ten times, a good performing company is expected to exceed its forecasted value and there are very few chances for them to be below the forecasted value. This forecast value accuracy is not a metric which will be provided by Evidently, and this is only to prove the performance of the model in order to explain the results of the thesis.

$$C = Calculated\ forecast\ value\ from\ the\ model \tag{6.5}$$

$$D = Actual\ sales\ happened\ in\ that\ time\ period \tag{6.6}$$

$$Forecast\ Value\ accuracy = C/D \tag{6.7}$$

- Visualization Benefit: The visualization impact cannot be monitored with the help of any metrics. The visualization can be interpreted in terms of the benefits it brings to analysing and improving the forecast process. The dashboards provide three different information related to Pipeline Quality, Pipeline hygiene and also acts as an output interface to the predictor and mathematical approach's results. The benefit of the visualization is not immediate and rather it takes a long term approach. Over time, these dashboards can be interpreted and could form an integral part of discussions between the sales teams and the management teams to improve the quality of the data in the pipeline and also improve the business standards of the company. When actions are appropriate, it will lead to improvements in the forecasting accuracy and classification accuracy, as these dashboards are a direct reflection of the Sales pipeline data being used.

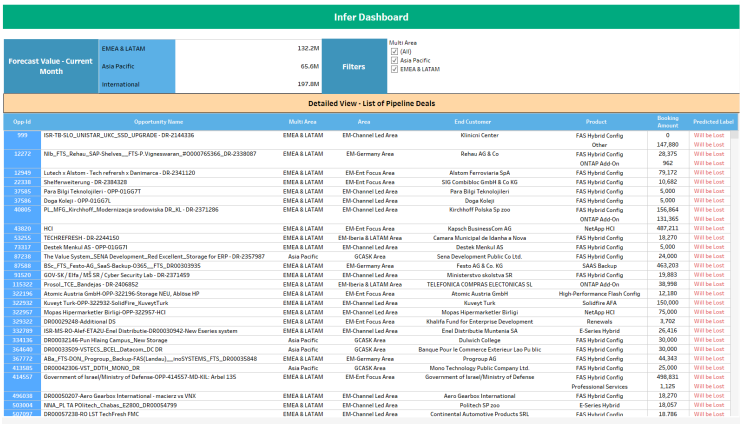

Figure 6.6: Predictor Output Dashboard

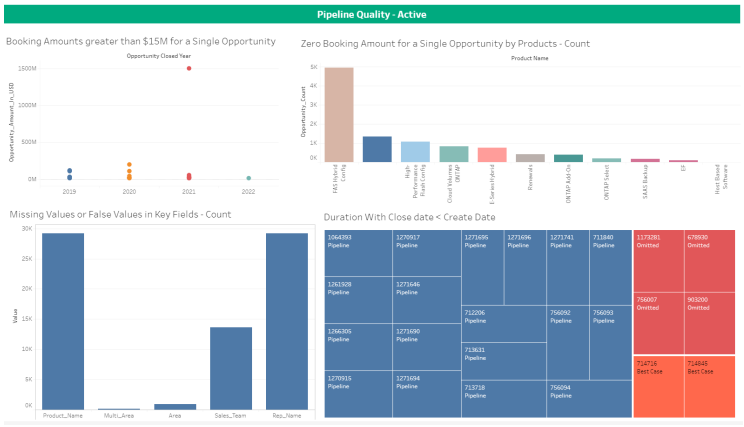

Figure 6.7: Pipeline Hygiene - Active

Figure 6.8: Pipeline Health - Past

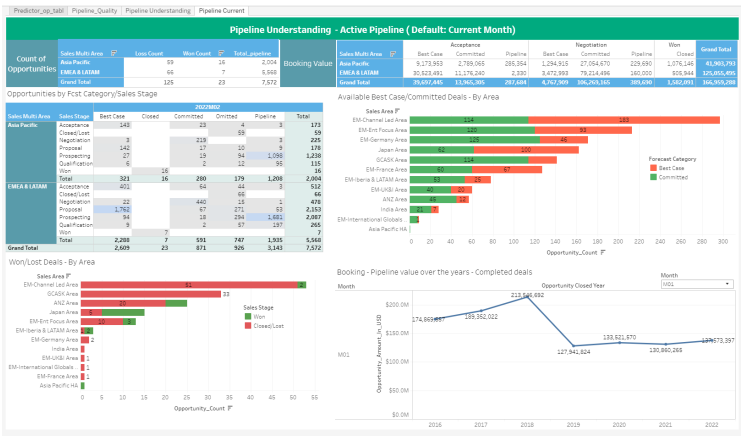

Figure 6.9: Pipeline Health - Active

7

Challenges of Machine Learning in B2B Sales Forecasting

With Machine learning on B2B Sales data, we came across some interesting challenges with respect to data handling and classifier performance. In this chapter, we try to list out the challenges faced and explain the approaches we took to fully or partially resolve them. Here, we focus more on B2B Sales Forecasting and the challenges faced over here are applicable to B2B based sales pipeline data. Most of the research in this area doesn't discuss the after effects of the machine learning model, but instead shed light on what kind of business challenges we can expect from the data. This was the main motivation to write this chapter.

7.1 Categorical Encoding Difficulties

The idea behind why we opted to encode the categorical values using CatBoost encoding is explained in section 4.3.3. Additionally, the motivation is also driven by the challenges we faced with categorical encoding of the data in the Active flow. The B2B sales pipeline data used for this thesis is highly categorical, and some features are highly cardinal as well, which were already explained in the previous sections. The B2B Sales pipeline data is constantly changing and although the features remain the same, their values can change over time. During the testing of the active flow, we actually tested two classifiers - namely RandomForest classifier and CatBoost classifier. Although CatBoost was selected in the end, we wanted to test the usage of Random Forest as well since it performed too close with the CatBoost. We designed two implementations of the active flow using the above two classifiers to monitor the performance. On running the two implementations, we found

that the RandomForest Classifier was failing. Upon investigation, we observed the presence of new values for some categories, which the classifier was unable to interpret as it has not seen it during training. The potential areas where we could identify changes in categorical values were

- Sales Team - There were new sales teams created, and some teams were split into two with newer names to suit the changes in the business model. This led to the active flow receiving these values, since the data for the active flow is new and fresh.
- Products - New portfolio of products or renaming the existing products. The effect of renaming the existing products was minimized by mapping the new names with the existing names in the data preprocessing stage.
- Areas and Regions - Changing the business model or upgrading it obviously has an impact on how the mapping of areas and regions are made. Like products, there were newer areas and some existing areas were either grouped into another one or split into different ones.

This kind of difficulties are more common in a B2B sales pipeline as companies evaluate and upgrade their business model and portfolios every year to be in line with the evolving market conditions and also to improve their business understanding. We resolved this issue by using the CatBoost encoding which has inbuilt capacity to account for unseen values by using its encoding methodology as explained in the section 4.3.4.The CatBoost classifier using the CatBoost encoding scheme (15) was able to successfully run and produce the expected results in spite of encountering new data. One can also wonder that to counter the effect of the splitting up of already existing feature values into newer values (for example: North Holland into Amsterdam and Amstelveen) could be offset by retraining the machine learning model and is not a categorical encoding problem. Yes, that is a possibility, but we do have some issues on the data side to successfully apply the method, which will be explained in the later sections. This problem of categorical encoding also gives rise to the problem of Data Drift.

7.2 Data Drift

Models which deal with streaming data experience the problem of data drift. The production models are faced with new data, and the distribution of the data can change over time due to a variety of factors. Data drift is the change in the distribution of the baseline

dataset on which the model was trained and the current real time production data. In our case, we evaluate the data drift with the help of Evidently AI's data drift reporting. As shown in the 5.4 we have a model performance classification block that monitors the production model on a scheduled basis and generates performance reports, and it also provides us with the Data drift report. Classification accuracy of Machine learning models generally have the ability to deteriorate over time with changes in data, which is termed as Model drift. Model drift is primarily caused due to two factors - Concept Drift and Data Drift. Concept drift is the phenomenon that happens when the meaning of what you are trying to predict changes. Concept drift is concerned with the changes in the prediction of the target variable.(29)(30)(31) For example, a corona prediction model designed to predict whether a person will have corona or not would have experienced concept drift. This is because, in the beginning only few symptoms of corona were known, and the model predicts based on those but with the increase in research we have identified newer symptoms that are unknown to the model. The concept has changed over time and if we don't retrain the model, then it is going to perform badly. The motivation to looking into this aspect of production based machine learning models stemmed from our model's drop in classification levels when connected to the active flow. We experienced a drop of 6% initially in our classification accuracy from our estimated classification accuracy of 77% on the test dataset of the training flow.

Upon investigation of the data drift report generated from the model by comparing the data from the training pipeline and the fresh incoming data from the Active pipeline, we formulated the following points which could have been producing the data drift,

- Week :- The week feature consists of one extra week for leap years. Our training data doesn't contain data for leap years. For leap years, we have 53 weeks compared to the normal 52 weeks. This means that there is one week of data that is unseen by the model, and the distribution of the week is not present in the training dataset.

- More Samples:- In some instances, the incoming data on the active flow had a higher distribution for particular features. For example, only five samples for a particular category in the training data, but there are more samples in the incoming data. This can lead to concept drift, where newer combinations are coming in which the model has not yet trained on.

- New Samples:- As stated in the previous section on the category encoding challenges, newer values can come in which are not a part of the training set, leading to changes

in distribution. This differs from the previous point because here we have entirely new values, whereas in the previous one, we have the values, but newer combinations could be potentially made.

The figures 7.1 and 7.2 shows some features and their interesting data drift charts.

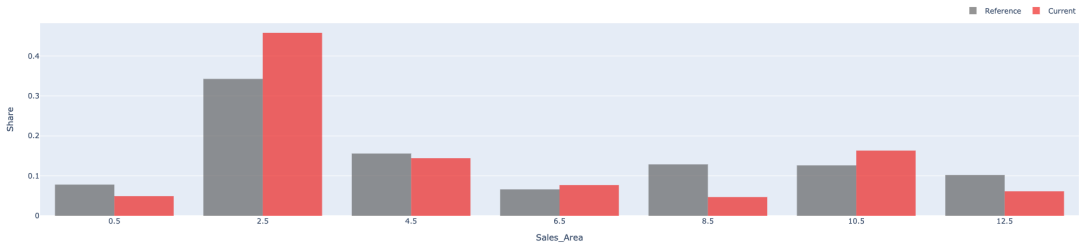

Figure 7.1: Drift - Sales Area

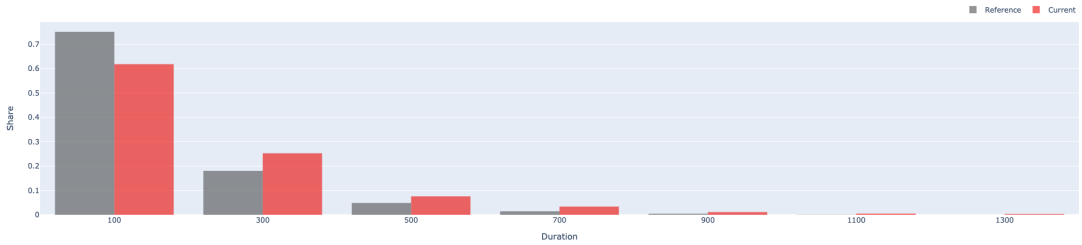

Figure 7.2: Drift - Duration

With that, we tried working around some features to improve the production model. For instance, we dropped the week level feature since it was a great cause, and we also dropped a similar feature which was representing the segmentation of customers as per the area in which they operate. With these, we were able to improve the model to predict at a 75% of accuracy, which is four points from the original state. We cannot classify it as a resolved issue because this is currently the case and this can change in the future as some other features can drift from the trained sample. This is mostly a problem when you use categorical data like the ones present in the B2B sales dataset. Some of the data challenges we had are explained in the next section.

7.3 Data Challenges

As the B2B business process keep changing over time, the data keeps changing as well. We list out potential data challenges which was interesting to understand when building the model.

- The ability of the entire data structure to change itself. Mostly B2B Companies group data based on the sales territory, customer segmentation and product grouping hierarchies. This is a periodic change in the companies with change in the business scheme of things. For example, today you have 59,000 records grouped under a geography like Europe. Tomorrow the company implements a new plan where due to business needs they are going to split the geography of Europe into Europe-West and Europe-East & Central. This would potentially mean the data structure is changing, and the machine learning models needs to be retrained and figured out again.
- Data Quality: As mentioned across the thesis, data quality is a major issue in formulating correct forecasts or predictions.
- Reinstatements of Data as per new structure takes longer time in a company as years of historical data needs to be aligned into the newer format, which potentially hinders the machine learning approach's capability to predict correctly. This is because with longer wait times, the model is still trained on the old set of data, which would give rise to the other problems mentioned in the previous sections of this chapter.

7. CHALLENGES OF MACHINE LEARNING IN B2B SALES FORECASTING

8

Limitations and Future work

The thesis gives an alternative approach to formulate the forecast in a B2B Sales setup. Although we were able to successfully devise a strategy to formulate the forecasts using the Sales pipeline data, there are some drawbacks which we don't fully overcome. This chapter explains in brief the limitations of the proposed framework and also talks about the future work which can be done to improve the forecasting process using machine learning.

8.1 Limitation

Every approach has its own advantages and limitations. It is impossible to design a framework or a working methodology without any limitations attached to it. In this regard, our framework has three specific limitations. Firstly, the model is seen as an alternative approach to use machine learning and arrive at a forecast which will challenge the conventional forecasting methodology. The problem stems from the adaptability of the sales professionals and the sales teams within the organization. We conducted an initial one on one discussions with a few members from different sales teams present in the organization to understand the likeability of this data driven approach. The outcome of the discussions was that although they were accepting the level of estimates the model produced, yet they were not so much in favour of considering it as a challenger. Organizations are also not fully interested to drive away from the conventional forecasting process or look at an alternative to challenge the forecast methodologies. Thus, we find very few related literature in this area when compared to sales forecasting in B2C domains where machine learning is becoming widely accepted.

The second limitation is the mathematical approach present in our framework. This works for most of the scenarios to churn out a stable forecast, but the mathematical

approach has no use if the sales value associated with an opportunity is entirely correct. This is an ideal scenario, but the mathematical approach can be discarded even if we have ninety percent of opportunities with correct values, as it will not increase or decrease the forecast value to induce the problems of under/over forecasting. The same data is used to formulate the conventional forecasts, but the sales professionals manually ensure that the correct sales value are recorded for each opportunity before using these opportunities for forecasting. This is also done on offline Excel files, and it's not corrected back in the CRM system from which we pull the data.

The third limitation is that at this moment we predict based on open and closed opportunities. Related work in this field also shows some approaches where they consider the opportunities which are in the other sales stages as shown in 2.1 and mark them as active and predict the same. When one looks up the historical sales pipeline data, There are very few opportunities which will be in the active state from the past. But this scenario is found in abundance with opportunities that are in the current or future sales pipeline. Although, our model learns the won/lost behaviour of the opportunities and uses them to predict the new deals, it's still a valid scenario to look into how to predict the different stages of the opportunities instead of just predicting the final stage which can help organizations to formulate other risk measures related to forecast and build more trust around the data driven model.

The last limitation which we would like to talk about is the ability of the external factors to influence the forecast processes. Although we don't see any drop in the sales behaviour of the organization due to pandemic, which helped to build a stable model, it is still not so easy to discard. The pandemic caused organizations to back out of their plans or downsize, leading to many significant cut down in revenue. Since B2B Sales is based on business to business, these factors will have a huge impact in estimating the forecast for a time period. Also, there are organization based issues (32) which can make a business back out of from its promise of buying, social and demographic factors like change of policies, governmental changes of an area etc. which can potentially impact the forecast. The data driven approach doesn't take into consideration factors like these while formulating the forecast, which makes it a less likely solution at times when these factors play a role. With this we provide an overview of the limitations and although there can be many more small limitations associated with the approach.

8.2 Idea - All stage B2B Predictor

As a future work, we would like to propose an idea before we jump into the conclusion part of the thesis. To approach the third limitation stated in ~~the final~~ section 8.1, we propose to take a different approach in terms of data storage, which can elevate the forecasting approaches. At the moment, each opportunity is represented with the means of a unique row in the dataset and the different stages of the opportunities are not recorded individually, but instead the sales professionals update the stage of the opportunity to the current stage. This leaves us with no ability to track the different stages the opportunity went through before finally going into the won/lost stage. As mentioned in the Background chapter of this paper, it is not a golden rule that the opportunity needs to go through every sales stage in order to get closed. It can directly go to the closed stage from the initial stage, or it can traverse some stages before it goes into the final stage. Keeping all records related to an opportunity pertaining to different sales stages is a mundane task and will result in the requirement of increased storage spaces and large volumes of data. So what we propose is to add another dimension to the data, where we have separate indicator features for each sales stage. A "Yes" will denote that the opportunity passed through this stage, and "No" will denote the opportunity did not go through that stage. With this we can identify the behaviour of the opportunity in terms of traversing the sales stage which will give a comprehensive picture of the progress an opportunity makes. This will also enhance the classification metric as the model can learn from the behaviour whether the opportunity will make a traversal across different stages or just abruptly enter into the omitted or lost stages. With this, we can also somewhat positively influence the final limitation mentioned in ~~the~~ section 8.1 which states the unaccountability of external factors. By noting the traversal of the stages, we can account for the buyer based factors such as backing out of a promise, internal problems etc as these deals don't traverse different stages, but they are just marked as omitted and are not considered for forecasting.

Although this idea seems like a change to the normal approach, the ability to learn about the opportunity's traversal of different stages will bring added benefits to the forecasting in a data driven approach.

9

Conclusion

Finally, we arrive at the last section of the thesis. With B2B Sales forecasting as the area of research for the thesis, we designed a model that consisted of three different components that together work to achieve the goal of the thesis, which was to use a data driven method to predict the won/lost nature of the opportunities and estimate a forecast value. The evaluations of the results from the model were promising and helps to provide an alternative for the conventional forecasting methodology followed in companies. Additionally, it gives the ability to make data driven decisions instead of having only the results from the conventional forecasting processes that are influenced by decisioning of the individual or the team. We list the research questions formulated for the thesis and state the conclusion we got from them.

1. RQ1.0 - Can a Sales Forecast formulated using Machine Learning challenge the Sales Forecast generated by the Knowledge of the sales team in terms of Quality?

   The model formed by using the three components mentioned provides an alternative data driven approach which can be used to challenge the conventional forecasting processes of the sales team to an extent. With the quality of the data we have at hand, we were able to produce a 77% classification accuracy on predicting the opportunities and the forecast value formulated using both the classifier and the mathematical approach was found to be 83% accurate to the actual sales happened in the same time period. With improvements in data quality as mentioned in the above chapters, the model can be improved even further.

   - Which Machine Learning Algorithm performs better with respect to the prediction?

From analysing the classification metrics we took to evaluate the different classifiers, we found out CatBoost was found to be the best model for carrying out the predictions.

- How expert's knowledge can be incorporated into modelling the data?

  The expert's knowledge was included in many sections of this model. It was included to carry out discussions to find out about the adaptability of the model. Expert's knowledge was instrumental in developing an understanding of the data. It also helped in feature selection, as we consider some features as part of our model which are not found to be important by the feature selection approaches. The knowledge was also instrumental in formulating the mathematical approach to achieve the forecast values and also in evaluating the results of the model. So the expert's knowledge was an integral part in bringing this model to life.

- How the model results can be evaluated?

  We use different classification metrics to evaluate our model such as the AUC-ROC curve, F1 Score and the classification accuracy. We also design a metric to calculate the accuracy of the mathematical approach, which are discussed in the experiment & evaluation of results chapter.

- What level of accuracy is accepted for the model to be deployed commercially?

  From related work, we could see higher accuracies obtained for the similar problems. Keeping the quality of the data at hand, we aimed for a 75% accurate forecast on predictions and the estimated forecast value to be at least 80% accurate. We are happy to see that the model's performance exceeded these accuracies.

- RQ2.0- How Data Drift plays a role in B2B Sales Data? Does it have a significant impact, and What can be done to minimize the changes in model performance, if any?

  We faced different challenges when coming up against the challenges of machine learning in B2B Sales forecasting. Difficulties in categorical encoding, Data challenges related to B2B Sales Data and data drift. We observed data drift on the active flow

of the model, which uses the new incoming data into the system as the input. The model performance deteriorated when handling new data from 77% to 71%. With Evidently AI, we formulated the data drift report, which helped us analyse which features contributed to this phenomenon. Based on the drift report, we made removed some features which helped improve model performance back to 75% in terms of classification accuracy. This data drift report was also monitored and formed a part of the active flow to identify if the model needs to be retrained or not and to keep a check on the distribution of the incoming data.

1. What are the KPI's with respect to Sales Pipeline/Predictor Output and does using Visualizations a good way to understand the Sales Pipeline?

   We initially wanted to look at formulating KPI's with which we designed this research question, but this changed during the course of the thesis. As there were multiple factors, designing KPI's for each of them would have resulted in an exodus of KPI's which would have not helped the scope of the thesis. The idea was altered to provide visualizations which can bring discussions about the Sales pipeline and also help improve the quality of the data flowing into the system. Also, visualizations were used to convey the results from the forecasting process. Thus, with this we formulated the visualizations to cover the three topics - Sales Hygiene, Sales Health and the output of the forecasting process. This provides an understanding of the sales pipeline and also helps to achieve a learning about the sales pipeline and provides a way to improve them.

   - What kind of Visualizations can be used for this process?

     Four separate dashboards namely Predictor Output dashboard, Pipeline Hygiene - Active, Pipeline Health - Active/ Past were created consisting of different types of visualizations to support the three areas we intended to cover. The individual visualizations present in the dashboards are discussed in chapter 5.

Results from this thesis can be used to add some significant research into the application of Machine learning techniques in the area of B2B Sales Forecasting. This also provides a way for organizations to apply this and start using and trusting machine learning in this area. While we have provided a work which provides acceptable results, this is still an area where a lot can be done. More work needs to be done by companies in terms of recording

different dimensions of data and improving the quality of the data in the Sales pipeline. Due to limited resources, we couldn't perform some machine learning side techniques to improve the classifier's performance like grid search to find optimal parameters as they were computationally expensive with increased data size. Also, with the evolving technologies, sales professionals should be encouraged to develop an understanding of machine learning and advanced techniques to enable researchers to explain about the benefits of them. This will help improve adaptability. Furthermore, as a conclusion, neural networks and deep learning models are not used in this area due to their issues with explainability. Work needs to be done to improve the explainability of these models, which can provide a lift to using data driven approaches in this area. Apart from these, machine learning still works as a good alternative to conventional forecasting techniques to formulate the forecasts with which we would like to conclude this thesis.

References

[1] Marko Bohanec, Mirjana Kljajic Borstnar, and Marko Robnik-Sikonja. **Integration of machine learning insights into organizational learning: A case of B2B sales forecasting**. In *28th Bled eConference: s#eWellBeing, Bled, Slovenia, June 7-10, 2015*, page 33, 2015. 13

[2] Junchi Yan, Min Gong, Changhua Sun, Jin Huang, and Stephen M. Chu. **Sales pipeline win propensity prediction: A regression approach**. In Remi Badonnel, Jin Xiao, Shingo Ata, Filip De Turck, Voicu Groza, and Carlos Raniery Paula dos Santos, editors, *IFIP/IEEE International Symposium on Integrated Network Management, IM 2015, Ottawa, ON, Canada, 11-15 May, 2015*, pages 854–857. IEEE, 2015. 14

[3] Stephen Mortensen, Michael Christison, BoChao Li, AiLun Zhu, and Rajkumar Venkatesan. **Predicting and Defining B2B Sales Success with Machine Learning**. In *2019 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–5, 2019. 14

[4] Lior Rokach and Oded Maimon. *Decision Trees*, **6**, pages 165–192. 01 2005. 14

[5] Leo Breiman. **Random Forests**. *Machine Learning*, **45**(1):5–32, 2001. 14

[6] Tianqi Chen and Carlos Guestrin. **XGBoost: A Scalable Tree Boosting System**. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016. 14

[7] Brendan Andrew Duncan and Charles Peter Elkan. **Probabilistic Modeling of a Sales Funnel to Prioritize Leads**. In Longbing Cao, Chengqi

Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1751–1758. ACM, 2015. 14

[8] Alireza Rezazadeh. **A Generalized Flow for B2B Sales Predictive Modeling: An Azure Machine Learning Approach**. *CoRR*, **abs/2002.01441**, 2020. 15

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3146–3154, 2017. 15

[10] Xin Xu Lei and Tang Venkat Rangan. **Hitting your number or not? A robust & intelligent sales forecast system**. In Jian-Yun Nie, Zoran Obradovic, Toyotaro Suzumura, Rumi Ghosh, Raghunath Nambiar, Chonggang Wang, Hui Zang, Ricardo Baeza-Yates, Xiaohua Hu, Jeremy Kepner, Alfredo Cuzzocrea, Jian Tang, and Masashi Toyoda, editors, *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 3613–3622. IEEE Computer Society, 2017. 15

[11] Tiemo Thiess, Oliver Müller, and Lorenzo Tonelli. **Design Principles for Explainable Sales Win-Propensity Prediction Systems**. In Norbert Gronau, Moreen Heine, Hanna Krasnova, and K. Poustcchi, editors, *Entwicklungen, Chancen und Herausforderungen der Digitalisierung: Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik, WI 2020, Potsdam, Germany, March 9-11, 2020. Zentrale Tracks*, pages 326–340. GITO Verlag, 2020. 15

[12] Scott M. Lundberg and Su-In Lee. **A Unified Approach to Interpreting Model Predictions**. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017. 16, 46

[13] Kedar Potdar, Taher S. Pardawala, and Chinmay D. Pai. **A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers**. *International Journal of Computer Applications*, **175**:7–9, 2017. 22

[14] Patricio Cerda and Gaël Varoquaux. **Encoding high-cardinality string categorical variables**. *CoRR*, **abs/1907.01860**, 2019. 22

[15] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. **CatBoost: unbiased boosting with categorical features**. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6639–6649, 2018. 23, 32, 52

[16] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. **CatBoost: gradient boosting with categorical features support**. *CoRR*, **abs/1810.11363**, 2018. 23, 32

[17] Essam Al Daoud. **Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset**. *International Journal of Computer and Information Engineering*, **13**(1):6 – 10, 2019. 23

[18] Girish Chandrashekar and Ferat Sahin. **A survey on feature selection methods**. *Comput. Electr. Eng.*, **40**(1):16–28, 2014. 24

[19] S. Visalakshi and V. Radha. **A literature review of feature selection techniques and applications: Review of feature selection in data mining**. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–6, 2014. 24

[20] Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. **Boruta - A System for Feature Selection**. *Fundam. Informaticae*, **101**(4):271–285, 2010. 24

[21] Daniela Gotseva Roumen Trifonov and Vasil Angelov. **Binary classification algorithms**. *International Journal of Development Research*, **07**, 11 2017. 30, 31

[22] Zhi-Hua Zhou. **Ensemble Learning**. In Stan Z. Li and Anil K. Jain, editors, *Encyclopedia of Biometrics, Second Edition*, pages 411–416. Springer US, 2015. 31

[23] Candice Bentéjac, Anna Csörgo, and Gonzalo Martínez-Muñoz. **A comparative analysis of gradient boosting algorithms**. *Artif. Intell. Rev.*, **54**(3):1937–1967, 2021. 31

[24] Daniel Berrar. *Cross-Validation*. 01 2018. 41

[25] **Catboost Cross validation**. 42

[26] **Evidently-AI**. 42

[27] Guy Handelma, Hong Kok, Ronil Chandra, Amir Razavi, Shiwei Huang, Mark Brooks, Marlon Lee, and Hamed Asadi. **Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods**. *American Journal of Roentgenology*, **212**:1–6, 10 2018. 43, 46

[28] Yangguang Liu, Yangming Zhou, Shiting Wen, and Chaogang Tang. **A Strategy on Selecting Performance Metrics for Classifier Evaluation**. *International Journal of Mobile Computing and Multimedia Communications*, **6**:20–35, 10 2014. 44

[29] Geoffrey I. Webb, Roy Hyde, Hong Cao, Hai-Long Nguyen, and François Petitjean. **Characterizing concept drift**. *Data Min. Knowl. Discov.*, **30**(4):964–994, 2016. 53

[30] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. **Learning under Concept Drift: A Review**. *CoRR*, **abs/2004.05785**, 2020. 53

[31] **Understanding and Handling Data and Concept Drift**. 53

[32] Donna Davis and John Mentzer. **Organizational factors in sales forecasting management**. *International Journal of Forecasting*, **23**:475–495, 07 2007. 58