Product Sales Opportunity Prediction

Harsha B M
CDS - DSBA
IISc
Bangalore, India
harshab@iisc.ac.in

Sunny Mattas
CDS - DSBA
IISc
Bangalore, India
sunnymattas@IISc.ac.in

Pranjal Gupta
CDS - DSBA
IISc
Bangalore, India
pranjalg@iisc.ac.in

Pranjal Gupta
CDS - DSBA
IISc
Bangalore, India
dineshkumar1@iisc.ac.in

***Abstract*— Product Sales Opportunity Prediction is a key challenge faced by businesses seeking to optimize their sales and marketing strategies. This paper presents an approach for data analysis and machine learning on sales pipeline data to understand the sales behavior, process gaps, and predict the probability of a sales opportunity being won or lost. This has the potential to enable timely insights for leadership team to better manage the sales pipeline and maximize revenue.**

Keywords—Sales pipeline, Machine Learning Classification, Sales Opportunity, Marketing, Bookings/Revenue, Data Science, Confusion Matrix

I. INTRODUCTION

For this paper, we look at a SaaS product of a Fortune 100 company. The product is sold through multiple sales and marketing channels like - partner, direct, events, advertising, webinars, website etc. These channels fill the sales funnel for the company. This is tracked through funnel stages – Lead, MQL, SQL, Opportunity and finally WON/LOST. The top and bottom of the funnel is managed by the company's marketing team and Sales team respectively. Our focus in the Opportunity to Close/Won phase of the pipeline.

II. REQUIREMENTS AND SETUP

A. Data

We were approved access to only 5 quarters of data for this work. The data is available in Snowflake cloud data warehouse.

B. Systems & Tools

The data model in Snowflake is complex. It is a STAR schema (Facts and Dimensions). The data volume for our analysis is also large (~1 million rows and ~120 columns). We also need access to SFDC to understand process.
Python and required ML packages were used in Jupyter notebooks.

C. Data Security

The data is company confidential. So, we must process and analyze data only using company's internal infrastructure and tools.

D. Business Understanding

The marketing and sales domain is new to us. We had to take guidance from Marketing & Sales leaders, sales managers, and analysts to understand the process and system. This is the key to put together an impactful and relevant work. that is actionable for the concerned teams.

For process understanding, we got access to Salesforce CRM which is the primary data source for Snowflake. This helped to see the process the action and make better sense of the data. This also helped us validate our understanding of the sales behavior and process gaps.

III. ACTIVITIES INVOLVED

Below is the set of activities performed for this work.

A. Data Preparation and Data Cleaning

- 6 quarters of sales pipeline data
- All open opportunities as of any given week of the quarter is the starting dataset
- Measurement of the outcome of these opportunities by the end of the quarter
- Data clean-up removing bad data

B. Exploratory Data Analysis

- To identify and understand the patterns and relationships in the data.
- To detect anomalies, outliers, and missing values in the data.
- To assess the quality and reliability of the data.
- To identify potential problems or limitations with the data.

C. Feature Engineering and Model Development

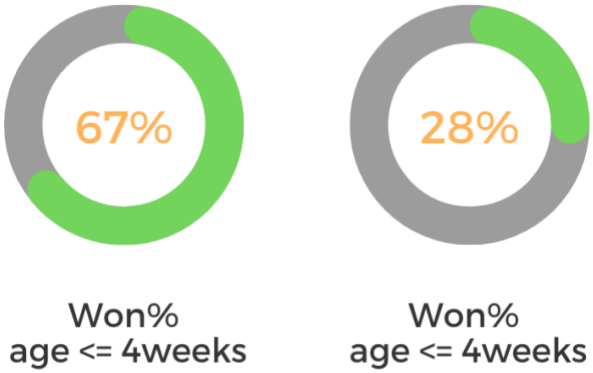
The primary focus is

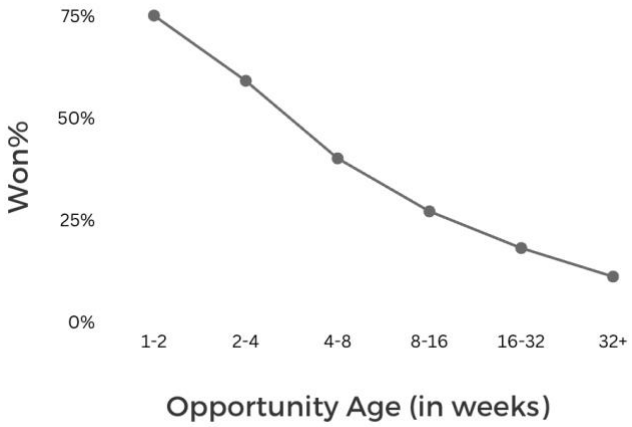
- Correlation Analysis
- Feature Selection
- Define binary Class Variable as opportunity Won Vs Not Won (Lost + Open at the end of the quarter)
- Divide dataset into training and testing datasets - 5 quarters of training data and most recent quarter of testing data
- Predict the probability of winning opportunity by the end of the quarter
- XGBOOST and Logistic Regression

IV. OUR FINDINGS

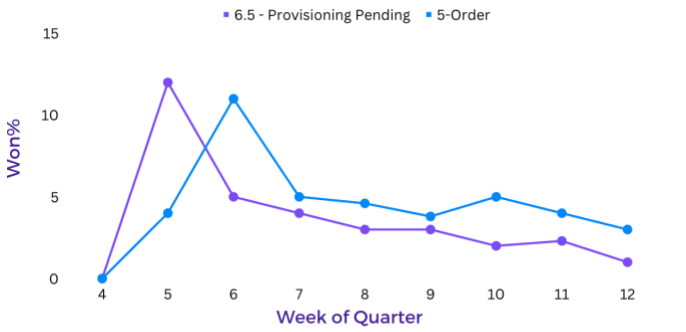
Below are our key findings from the data analysis.

*A. There is **2.5X** higher probability for an opportunity to be WON if closed in less than 4 weeks*





B. Probability of an opportunity to be WON decreases if it stays in the same stage (especially after stage 4-negotiate) longer than ~2 weeks

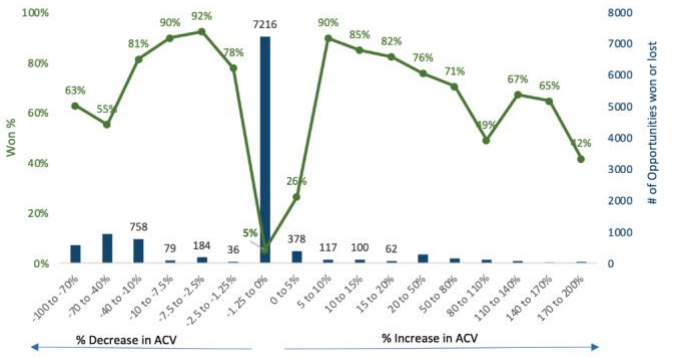


Won % for opportunities open in respective stage as of week 3

C. Active engagement with customers results in a Win-Win Scenario

a) On an average, 85% opportunities are won if ACV was either increased or decreased within a range of 5-40%.

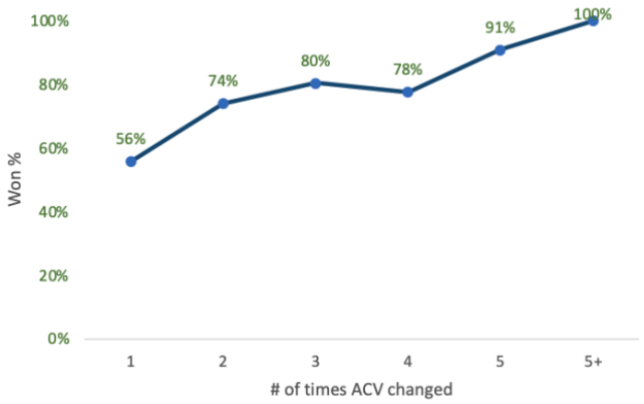
b) Only 5% of the opportunities are won when ACV remained unchanged.



[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

c) An opportunity has 31% more chance of winning when it undergoes more than 2 negotiations during its lifecycle



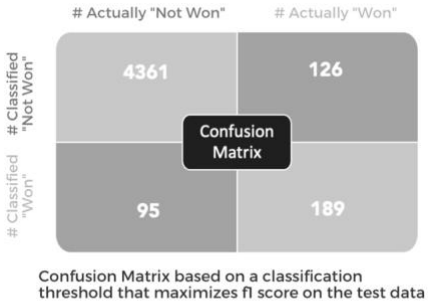
D. EXISTING base matters while we pursue NEW logos

a) Opportunities for EXISTING customers has 2.5X higher win-rate compared to NEW customers

b) Opportunities with "Product" as part of ELA has 2X higher win-rate than those where it was sold a-la carte

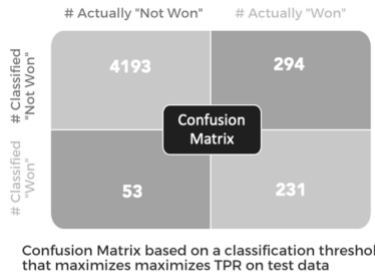
V. Model Results

A. We looked at Confusion matrix for the quality of our model.



Confusion Matrix based on a classification threshold that maximizes f1 score on the test data

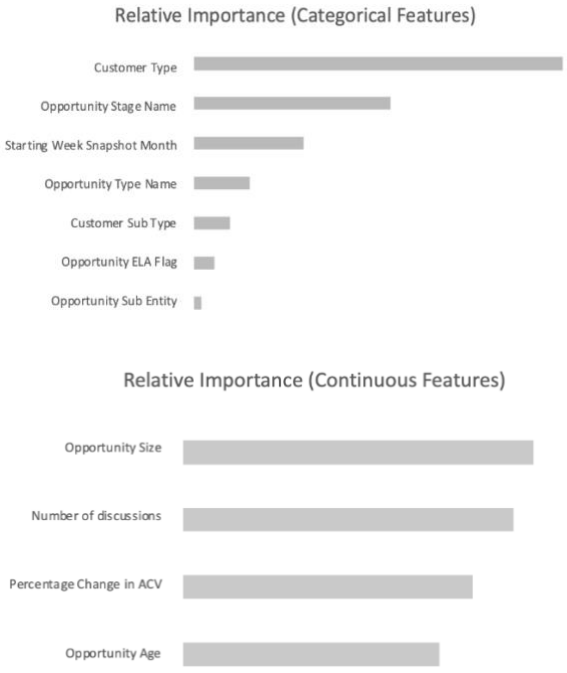
Model Accuracy Score 95.4%



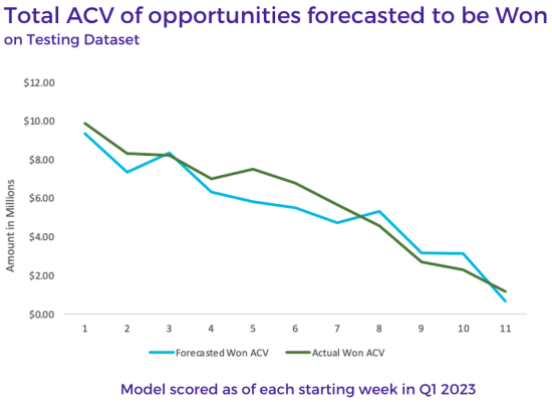
Confusion Matrix based on a classification threshold that maximizes maximizes TPR on test data

Model Accuracy Score 92.7%

B. Some key features for the model



C. The classification model performed very well on the test data



Acknowledgment

We would like to acknowledge the guidance we received from Prof. Sashikumaar Ganeshan, Thivin and Divij. They helped us with concepts required for this analytics work. We also thank the Sales Managers, SFDC Admins and Data Governance teams for their support. Their inputs on the process and data were very helpful to validate our understanding and findings.