MSc. in Data Analytics and Design Thinking for Business

DAS 601 - Basic Machine Learning and Artificial Intelligence for Creative Business Analysis

Final Project
On
An Explainable AI Study on Identifying Factors Affecting Customer Churn of A Telecommunication Company

Submitted By

Piyal Dey

ID: 233001861

Department of Digitalization, Innovation and Entrepreneurship

School of Business Administration

Submitted To

Musabbir Hasan Sammak

Visiting Lecturer

July 2024

East Delta University (EDU)

Noman Society, East Nasirabad, Khulshi, Chattogram: 4209

Table of Contents

ABSTRACT

Customer churn prediction is a critical challenge for telecommunications companies, as retaining customers is significantly more cost-effective than acquiring new ones. This study aims to develop an accurate predictive model for customer churn and to identify the key factors influencing customer decisions on leaving the service. The research applies a comprehensive feature selection process, evaluating Principal Component Analysis (PCA), filter methods (correlation, chi-square test, and mutual information), Recursive Feature Elimination (RFE), and LASSO regression, to refine the dataset and enhance model performance.

Five machine learning models Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Artificial Neural Network (ANN) were trained on datasets prepared by feature selection methods. Among these methods, models trained on features selected by the filter method displayed superior generalization capabilities. XGBoost and Random Forest appeared to be the top models based on performance. Overall, XGBoost slightly outperformed Random Forest after hyperparameter tuning, achieving a train accuracy of 92% and a test accuracy of 89%.

To serve the purpose of the study, SHAP (Shapley Additive Explanations) was employed, which revealed that two-year contracts, premium tech support, fiber optic internet type, and online security features significantly reduce churn likelihood. In contrast, month-to-month contracts, dissatisfaction, a high number of dependents, and credit card payment methods were identified as key drivers of churn. These insights can offer actionable recommendations for the telecommunications company to optimize customer retention strategies.

CHAPTER 1
INTRODUCTION

In today's competitive business world, customer relationship management (CRM) is a crucial strategy for every business. Many businesses explicitly focus on CRM in initiating, managing and strengthening the customer relationship for customer acquisition and retention. In the industry of telecommunication, businesses face extreme competition along with behavioral change in customer which results in churn. Churning refer to the scenario when customers closes subscription of a company and switches to another service provider. Factors like low customer satisfaction, new product launches, changes in regulations, changes in subscription model, etc. can be the reason of customer churn. Failure to proactively identify the factors that negatively affects customer satisfaction level and build strategies to retain customers will result in financial and reputational loss.

Machine learning has been considered as the most powerful tool to predict customer churn by many scholars. Over the last decade, many researchers have already proposed and tested several machine learning models to predict customer churn accurately **[1, 2, 3]**. Most popular machine learning models, for instance Artificial Neural Networks, Decision Trees learning, Regression Analysis, Logistic Regression, Support Vector Machines were implemented to solve this churn prediction problem. In general, the main focus of the studies was on improving the accuracy of the prediction.

However, identifying the underlying factors that impacts churn is equally important to build and iterate over strategies for customer retention. Previous studies have emphasized the need of identifying the factors that affect customers to churn along with the predictions. For better decision making, models' explanation behind its prediction is crucial. This can assist the telecom companies to implement strategies based on targeted factors. With the advancement of artificial intelligence, Explainable AI (XAI) models can offer valuable insights about the underlying factors driving the predictions and help business develop strategies based on these factors impact.

One major challenge is the need for models that not only predict churn accurately but also provide explanations for their predictions, facilitating better decision-making for business

managers. This need for explainability in AI models is critical in the telecommunications industry, where understanding the reasons behind customer churn can lead to more targeted and effective retention strategies. This research aims to address this gap by integrating explainable AI techniques into churn prediction models, thereby enhancing both the accuracy and interpretability of these models.

Despite the advancements in churn prediction models, there remain significant gaps in the interpretability of these models. Most existing research has focused on improving predictive accuracy without adequately addressing the need for model transparency. This lack of transparency can hinder the ability of telecom companies to understand customer behavior and effectively mitigate churn. By leveraging Explainable AI, this research aims to fill this gap by providing interpretable models that elucidate the factors leading to customer churn. This approach will empower telecom companies to devise more targeted retention strategies, thereby reducing churn rates and enhancing customer satisfaction.

The primary objective of this research is to develop an explainable machine learning model to predict customer churn in the telecom industry. After that, the aim is to identify and analyze the key factors influencing customer churn. Develop a machine learning model that not only predicts churn but also provides clear explanations for its predictions.

This research is expected to optimize model performance and identify key factors behind customer churn. Identification of these factors will allow telecommunication companies to develop and design customer retention strategies effectively.
The study is arranged into five chapters as follows:

- Chapter 1: This chapter introduces the research objectives, background, and significance of the study.
- Chapter 2: This chapter explains the methodology used in the research for dataset collection, data exploration, data pre-processing, and model implementation.
- Chapter 3: This chapter demonstrates the results from the analysis with interpretations.
- Chapter 4: This chapter discusses on the results from the analysis with interpretations.
- Chapter 5: This chapter outlines the conclusion and further work needs of the study.

CHAPTER 2
METHODOLOGY

2.1 Dataset

The public dataset is completely available on the Maven Analytics website storing all available datasets for analysis in the Data Playground. The specific telecom customer churn dataset at hand can be obtained in this link below: https://www.mavenanalytics.io/blog/maven-churn-challenge. The Customer Churn table contains information on all 7,043 customers from a Telecommunications company in California in Q2 2022. Each record represents one customer, and contains details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned), and more.

Field	**Description**
CustomerID	A unique ID that identifies each customer
Gender	The customer's gender: Male, Female
Age	The customer's current age, in years, at the time the fiscal quarter ended (Q2 2022)
Married	Indicates if the customer is married: Yes, No
Number of Dependents	Indicates the number of dependents that live with the customer (dependents could be children, parents, grandparents, etc.)
City	The city of the customer's primary residence in California
Zip Code	The zip code of the customer's primary residence
Latitude	The latitude of the customer's primary residence
Longitude	The longitude of the customer's primary residence
Number of Referrals	Indicates the number of times the customer has referred a friend or family member to this company to date
Tenure in Months	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above
Offer	Identifies the last marketing offer that the customer accepted: None, Offer A, Offer B, Offer C, Offer D, Offer E
Phone Service	Indicates if the customer subscribes to home phone service with the company: Yes, No
Avg Monthly Long Distance Charges	Indicates the customer's average long distance charges, calculated to the end of the quarter specified above (if the customer is not subscribed to home phone service, this will be 0)
Multiple Lines	Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No (if the customer is not subscribed to home phone service, this will be No)
Internet Service	Indicates if the customer subscribes to Internet service with the company: Yes, No

Internet Type	Indicates the customer's type of internet connection: DSL, Fiber Optic, Cable (if the customer is not subscribed to internet service, this will be None)
Avg Monthly GB Download	Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above (if the customer is not subscribed to internet service, this will be 0)
Online Security	Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Device Protection Plan	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
Premium Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming TV	Indicates if the customer uses their Internet service to stream television programing from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
Streaming Music	Indicates if the customer uses their Internet service to stream music from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to internet service, this will be No)
Unlimited Data	Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No (if the customer is not subscribed to internet service, this will be No)
Contract	Indicates the customer's current contract type: Month-to-Month, One Year, Two Year
Paperless Billing	Indicates if the customer has chosen paperless billing: Yes, No
Payment Method	Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
Monthly Charge	Indicates the customer's current total monthly charge for all their services from the company
Total Charges	Indicates the customer's total charges, calculated to the end of the quarter specified above
Total Refunds	Indicates the customer's total refunds, calculated to the end of the quarter specified above
Total Extra Data Charges	Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above

Total Long Distance Charges	Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above
Total Revenue	Indicates the company's total revenue from this customer, calculated to the end of the quarter specified above (Total Charges - Total Refurnds + Total Extra Data Charges + Total Lond Distance Charges)
Customer Status	Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined
Churn Category	A high-level category for the customer's reason for churning, which is asked when they leave the company: Attitude, Competitor, Dissatisfaction, Other, Price (directly related to Churn Reason)
Churn Reason	A customer's specific reason for leaving the company, which is asked when they leave the company (directly related to Churn Category)

2.2 Feature Engineering

There were several features that were skewed, imbalanced and in inappropriate data types for training machine learning model. To ensure optimization of model performance the features of the dataset were processed with the following feature engineering.

Zip Code Frequency Encoding: To capture geographic patterns, the Zip_Code_freq feature was created by mapping the frequency of each zip code present in the training data.

Age Binning: The Age feature was in numerical format, to better use age feature, it was divided into equal categorical bins using equal-width binning. This created labels that represent distinct age ranges (Age_bin).

Binary Encoding: Categorical variables having bool data was encoded into binary form so that they can be utilized in prediction model.

One-Hot Encoding: Another set of categorical features including Offer, Contract, Gender, and more were transformed into binary vectors using one-hot encoding method.

Scaling: All the features were then transformed into a specific range using standard scaler so that machine learning model can generalize well.

SMOTE: Initially the target variable was highly imbalanced, which will have an adverse effect on model training. To balance the target variable, SMOTE was implemented and later on this SMOTE transformed data frame will be used for model training.

Feature Selection Techniques

For feature selection, the study trained model on selected feature by 4 different method. These are, features selected by Principal component analysis (PCA), filter method (using correlation, chi-square test and mutual information), wrapped method using recursive feature elimination (RFE) and finally embedded method (using LASSO regression). 4 data frames were created based on the features selected by these methods. Then, machine learning model, Logistic regression, Random forest, SVM, ANN and XGBOOST were trained on these data frames. Based on the model performance, one method of feature selection is decided for further hypertuning and prediction on test data. The performance of these models will be discussed further in the next chapter.

Principal Component Analysis (PCA): PCA was applied to reduce the dimensionality of the dataset while retaining features that has most significant variance. This technique is a popular method for feature selection where it optimizes the prediction power even after reducing dimensionality.

Feature Filtering – Using Correlation, Chi-Square Test and Mutual Information:

At first, highly correlated features were identified and removed from the data frame to lessen the multicollinearity. This step ensures that redundant information does not bias the model. Then, chi-square test was used for assessing the independence between each feature and the target variable (Churn). Features having high chi-square scores indicates strong relationship with churn, were selected for further analysis. Finally, Mutual information was measured for emphasizing non-linear relationships with the target variable (Churn). Then after filtering features in three way, only the features that were selected by all the tests were selected and used for model training.

Feature Wrapping - Recursive Feature Elimination (RFE):

Another method of features selection, RFE which iteratively removes features based on their importance in the model by a random forest classifier. This method systematically identifies and retains the most influential features which enhances the model's robustness and interpretability.

Feature Embedding - LASSO (L1 Regularization):

Another approach to eliminate less significant feature is by penalizing less significant features. LASSO is applied on L1 regularization to shrink less significant features to zero and effectively select features with non-zero coefficients.

The effectiveness of each feature selection method was evaluated by training machine learning models on datasets refined by these techniques. Comparative analysis of model performance provided insights into the impact of feature engineering on predictive outcomes, supporting informed decision-making in customer retention strategies.

2.3 Model Selection

Initially 5 classification models were selected for predicting churn predictions. These are as follows:

Logistic Regression: Logistic Regression is a linear model used for classification tasks. It estimates the probability of a binary response based on one or more features. It provides a probabilistic framework that helps in interpreting the impact of each feature on churn prediction.

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. Its ability to handle a large number of features makes it a vigorous choice for churn prediction.

XGBoost: Extreme Gradient Boosting is a powerful ensemble technique that builds multiple weak learners (decision trees) sequentially. It continuously focuses on improving the model by reducing errors of the previous models. Its strength lies in its regularization capabilities, which prevents overfitting and overall predictive performance.

Support Vector Machine (SVM): SVM is a supervised learning algorithm that identifies the optimal hyperplane which maximizes the margin between the classes in the feature space. The SVM model was trained on the selected features from various feature selection techniques. It is particularly effective in high-dimensional spaces and with clear margin separation.

Artificial Neural Network: Neural Networks are computational models inspired by the human brain, consisting of interconnected units (neurons) that process information in a layered

structure. Its captures complex non-linear relationships in the data which makes it a versatile tool for classification prediction tasks.

2.4 Model Evaluation

Most studies using AUC and Accuracy to evaluate their model in CCP. Thus, this study used the confusion matrix to evaluate the implemented model and AUC metric. The accuracy, Recall, F1-Score, and precision can be obtained from the confusion matrix as shown in Table

Table 1: Confusion Matrix

		Predicted	
Actual		Churn	Not Churn
	Churn	TP	FN
	Not Churn	FP	TN

These metrics are suitable for analyzing any model built for classification problem. The following metrics have been thoroughly studied to explain the performance of described model:

1. A customer who is churning (positive) and classified as churn (positive) is called “True Positive (TP)”.
2. A customer who is not churning (negative) and classified as not churning (negative) is called “True Negative (TN)”.
3. A customer who is not churning (negative) but classified as churn (positive) is called “False Positive (FP)”.
4. A customer who is churning (positive) but classified as not churning (negative) called “False Negative (FN)”.

Accuracy: Accuracy is the percentage of the total number of predictions that were correctly classified. The metric is calculated from the equation

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}.$$

Precision: Precision is the proportion of the predicted true positive cases and is calculated from the equation

$$Precision = \frac{TP}{TP + FP}.$$

Recall: Recall is the proportion of positive cases that were correctly identified by the model. It focuses on the number of False Negative thrown into a prediction mixture. The recall is also known as sensitivity or true positive rate, and it is calculated as the following:

$$Recall = \frac{TP}{TP + FN}.$$

F1-Score: Precision or recall alone cannot describe the efficiency of a classier since good performance in one of those classes does not suggest good performance on the other class. For this reason, F1-Score, is considered as the most significant metric for evaluating classier performance. It is defined as the harmonic mean of precision and recall.

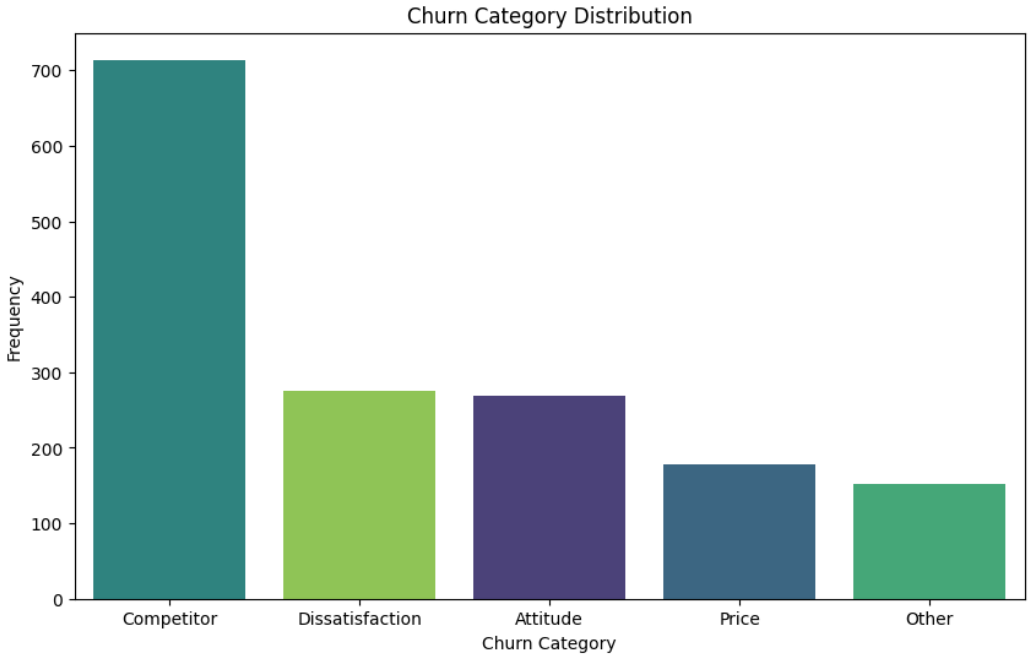
$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

The AUC evaluation metric is also used to measure the efficiency and performance of a binary Classifier. The AUC provides a more powerful evaluation metric than other evaluation metrics, which measures a supervised classification's overall performance by considering all potential cut-off points on the receiver's operating features curve.

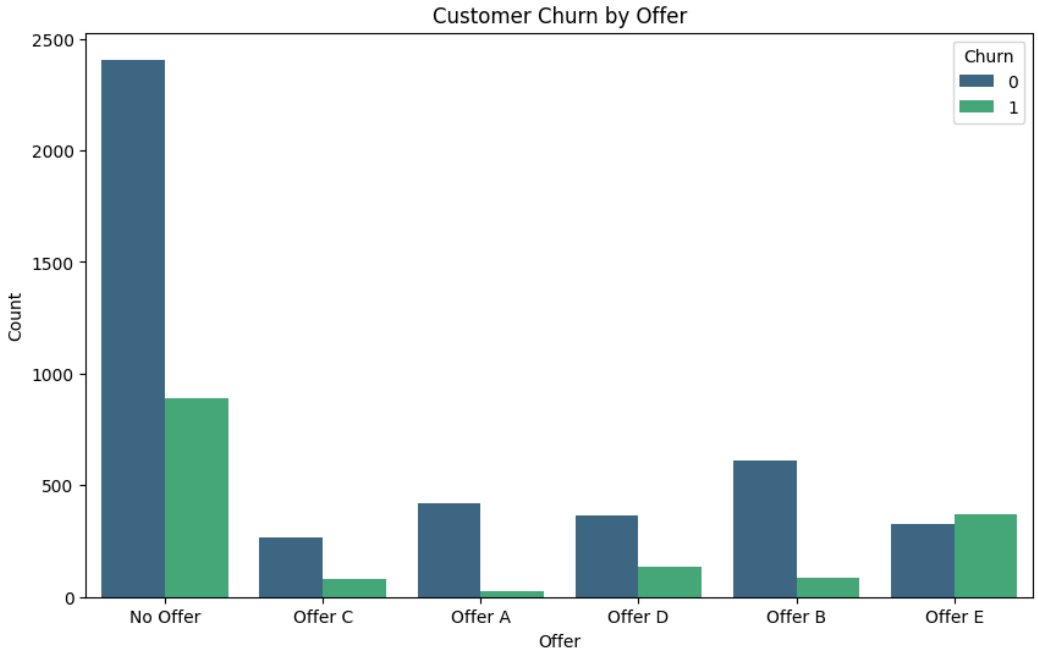
CHAPTER 3
RESULTS

3.1 Data

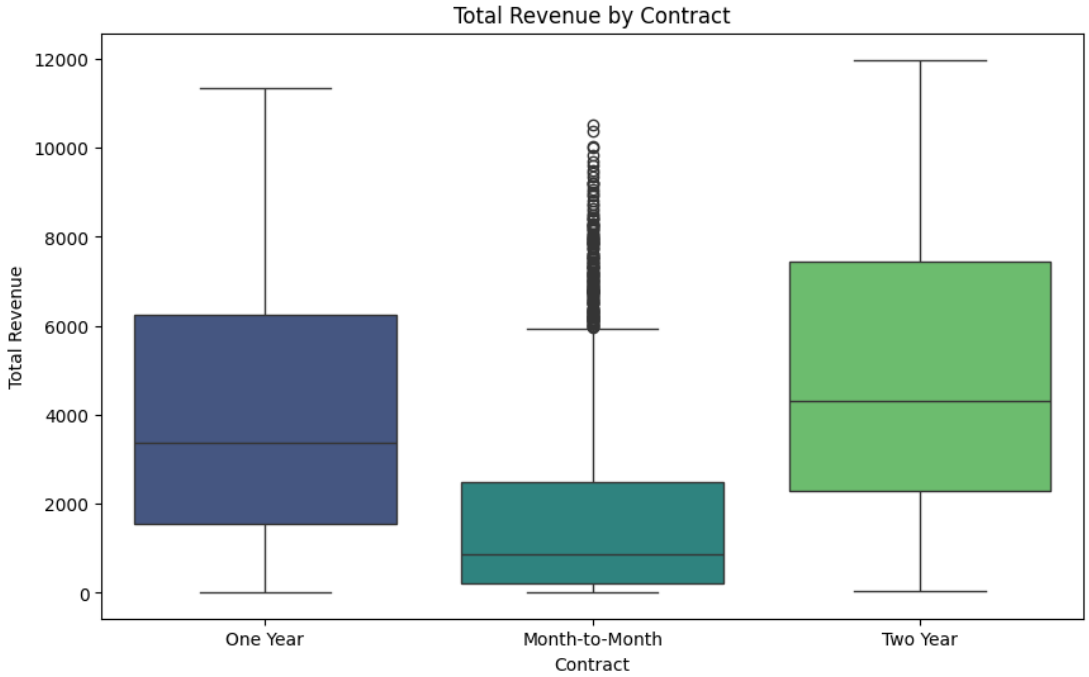
During the exploratory data analysis, many patterns on the data was discovered. Those are described below:



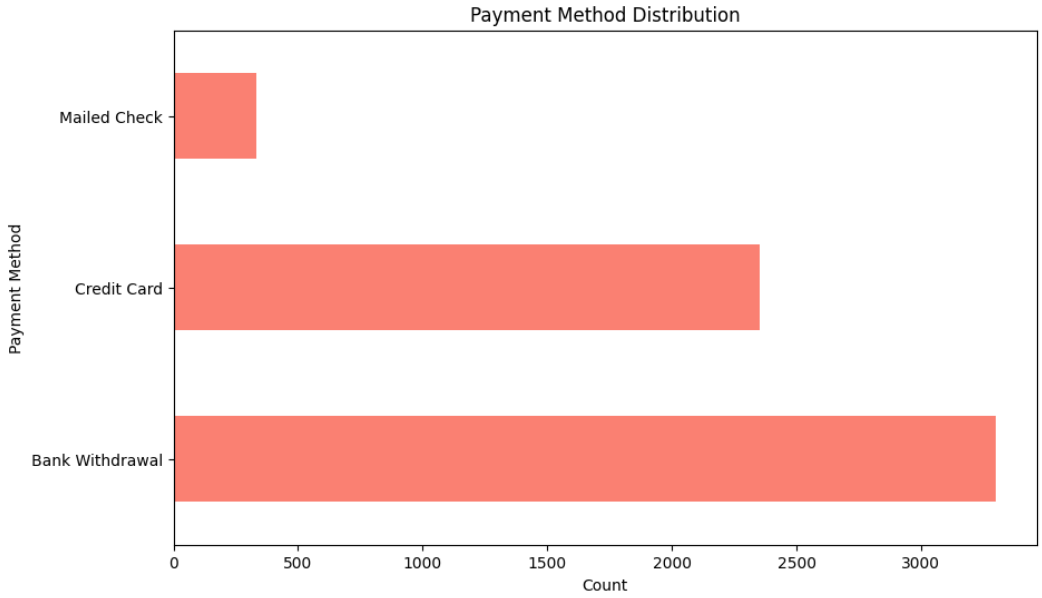
The churn category distribution indicated majority of the customers were leaving the service of the company due to better service availability by its competitors. This was further analyzed in the feature selection section on whether it had any significant impact on the churn. The filtered method identified both of these features significant for churn prediction.



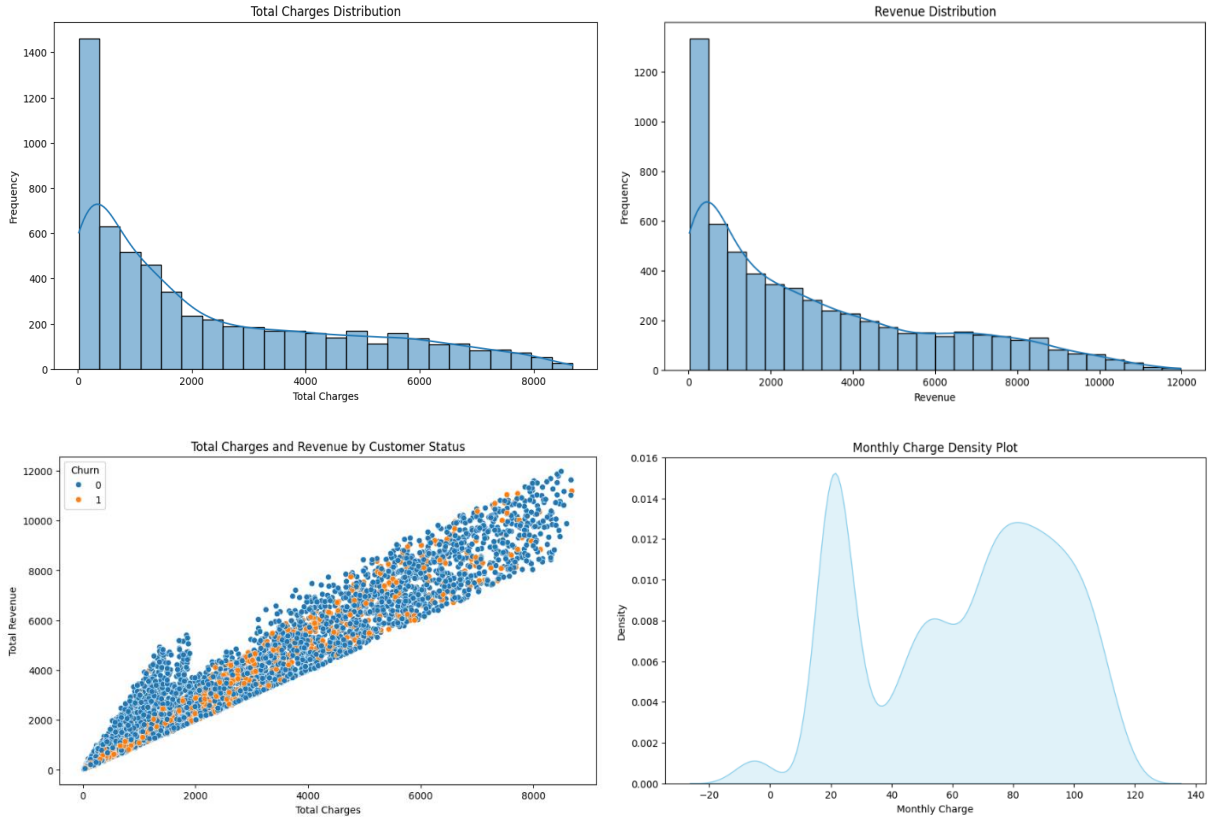
Another multi categorical feature was ‘Offer’ where it was observed that customer subscribed to no offer plans had high tendency to churn than other subscribed customers. This may also indicate the factor importance of churn. This was further analyzed on feature selection process to identify whether any particular offer subscription status has any impact on the customer churning. However, the filtered method didn’t consider offers as an impactful feature for churn prediction.



To dug deep into which contract types generate more revenues for the company, it was discovered that customers subscribed to 2-year contract generates higher revenue for the company. However, its monthly contract is the lowest income generating contract type and had severe number of outliers, which indicates that people are not satisfied with the product and service, and the customers of 1-2-year contracts will most likely to churn after their contract expires. The feature selection process identified month-to-month and two-year contracts highly significant for predicting customer churn status.

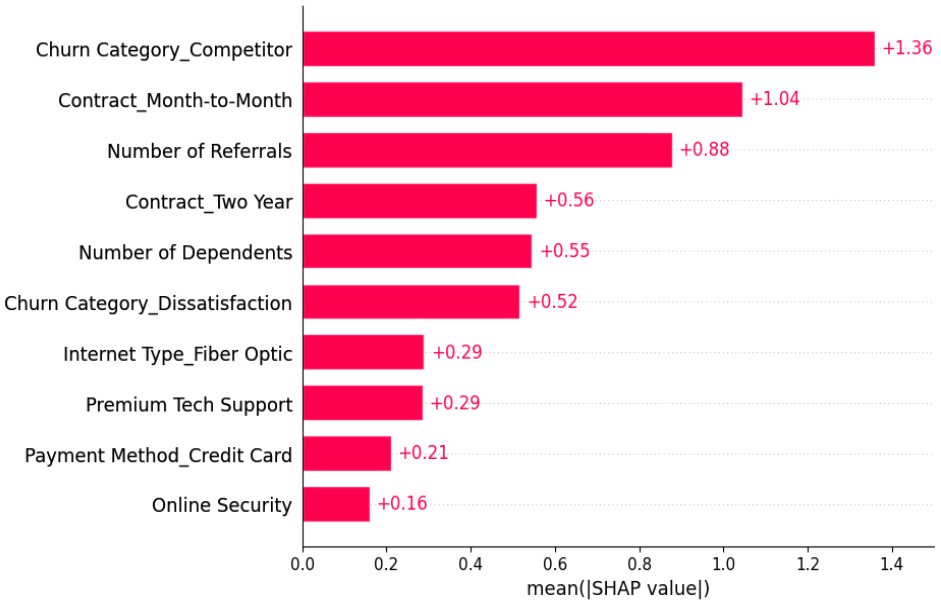


Initially, based on the visualization it was observed that most of the customers use bank withdrawal as a method for payment. However, feature selection process signifies only credit card payment feature to be more impactful on churn prediction.



The total and monthly charges, and revenue features showed high skewness which can adversely impact on the model performance. To handle this skewness, these features were

scaled in a consistent range using standard scaler. However, these features were not identified significantly impactful on customer churn by feature selection process.



During model evaluation on training set, different features were selected from PCA method, filtered method, wrapper method and embedded method. Features selected by using Filtered method performed generally better than other methods across all models. While other methods were overfitting even after intensive feature engineering, filter method was able to generalize well on unseen data. The comparative model performance is shown in the following section.

3.2 Feature Selection on Model Performance

Model performance on features selected by PCA

Metric	Logistic Regression	Random Forest	XGBoost	SVM	ANN
Precision (0)	0.90	1.00	1.00	0.90	1.00
Recall (0)	0.84	0.99	1.00	0.84	1.00
F1-score (0)	0.87	1.00	1.00	0.87	1.00
Support (0)	893.00	893.00	893.00	893.00	893.00
Precision (1)	0.84	0.99	1.00	0.85	1.00
Recall (1)	0.90	1.00	1.00	0.91	1.00
F1-score (1)	0.87	1.00	1.00	0.88	1.00
Support (1)	866.00	866.00	866.00	866.00	866.00
Accuracy	0.87	1.00	1.00	0.87	1.00
Macro Avg	0.87	1.00	1.00	0.88	1.00
Weighted Avg	0.87	1.00	1.00	0.87	1.00

Model performance on features selected by Filtered Method

Metric	Logistic Regression	Random Forest	XGBoost	SVM	ANN
Precision (0)	0.90	0.91	0.91	0.90	0.90
Recall (0)	0.86	0.94	0.94	0.86	0.91
F1-score (0)	0.88	0.92	0.92	0.88	0.91
Support (0)	893.00	893.00	893.00	893.00	893.00
Precision (1)	0.87	0.93	0.94	0.86	0.91
Recall (1)	0.90	0.90	0.90	0.90	0.89
F1-score (1)	0.88	**0.92**	**0.92**	0.88	0.90
Support (1)	866.00	866.00	866.00	866.00	866.00
Accuracy	0.88	**0.92**	**0.92**	0.88	0.90
Macro Avg	0.88	0.92	0.92	0.88	0.90
Weighted Avg	0.88	0.92	0.92	0.88	0.90

Model performance on features selected by Wrapper and Embedded Method

Metric	Logistic Regression	Random Forest	XGBoost	SVM	ANN
Precision (0)	1.00	1.00	1.00	1.00	1.00
Recall (0)	1.00	1.00	1.00	1.00	1.00
F1-score (0)	1.00	1.00	1.00	1.00	1.00
Support (0)	893.00	893.00	893.00	893.00	893.00
Precision (1)	1.00	1.00	1.00	1.00	1.00
Recall (1)	1.00	1.00	1.00	1.00	1.00
F1-score (1)	1.00	1.00	1.00	1.00	1.00
Support (1)	866.00	866.00	866.00	866.00	866.00
Accuracy	1.00	1.00	1.00	1.00	1.00
Macro Avg	1.00	1.00	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00	1.00	1.00

The models were trained on the 4 datasets separately with selected features by different approaches and evaluated with 5-fold evaluation metrics on validation set. The model's performance on validation set is shown above. Models performance on wrapped method and embedded method shown perfect accuracy and every other evaluation metrics, which seems unrealistic and showing signs of overfitting. The data frame of features selected by PCA also showed signs of overfitting. Only logistic regression and SVM model shown some degree of

generalization on prediction. Data frame created by filtered method has shown more degree of generalization on unseen data across all model's performance compared to other feature selection method. Random forest and XGBoost had shown 92% of F1-Score and Accuracy, higher than the other models. Therefore, these two models were selected as candidate model and further hyper tuned to train on best parameters.

3.3 Model Evaluation

Best Parameters for Random Forest Model obtaining best Cross-Validation Score 0.9171:

bootstrap: False, max_depth: 20, min_samples_leaf: 1, min_samples_split: 10, n_estimators: 200

Best Parameters for XGBoost model obtaining best Cross-Validation Score 0.9223:

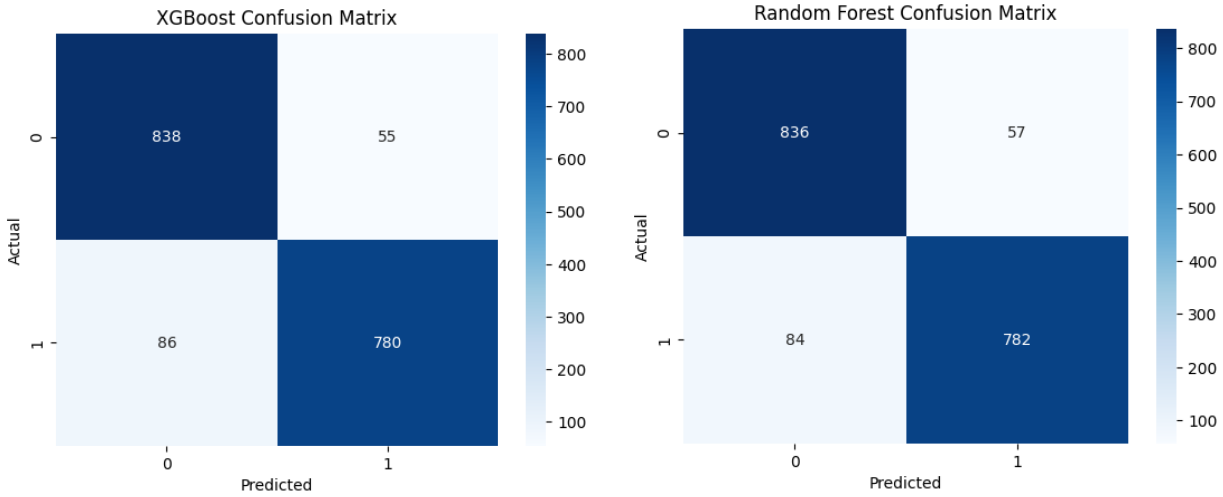
colsample_bytree: 0.8, learning_rate: 0.2, max_depth: 3, n_estimators: 200, subsample: 1.0

Model performance after Hypertuning

Metric	Best Random Forest	Best XGBoost
Precision (0)	0.91	0.91
Recall (0)	0.94	0.94
F1-score (0)	0.92	0.92
Support (0)	893.00	893.00
Precision (1)	0.93	0.93
Recall (1)	0.90	0.90
F1-score (1)	0.92	0.92
Support (1)	866.00	866.00
Accuracy	0.92	0.92
Macro Avg	0.92	0.92
Weighted Avg	0.92	0.92

After hypertuning, both of the models were trained on optimized parameters. The XGBoost model has performed slightly better than random forest. All the metric score was close to each other and showed similar performance. The confusion matrix also presented similar number of false positive and false negative. Overall, both of the models were similar in terms of

performance, however, XGBoost was slight ahead of random forest. Therefore, XGBoost was selected for run prediction of test dataset.

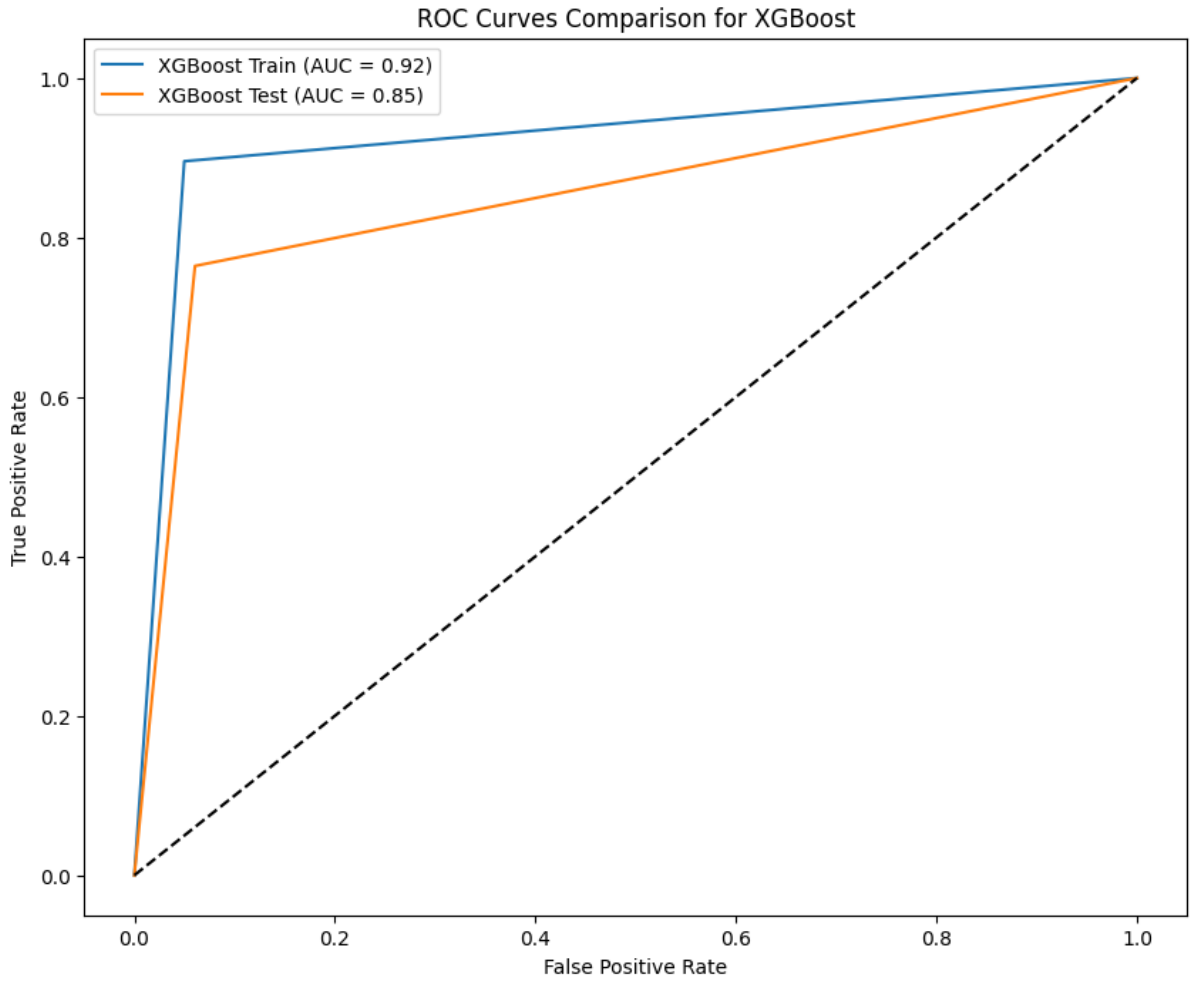


Here is the model performance of XGBoost on train and test set,

Model performance on Train and Test Set

Metric	Train Result	Test Result
Precision (0)	0.90	0.92
Recall (0)	0.95	0.94
F1-score (0)	0.92	0.93
Support (0)	4397.00	777.00
Precision (1)	0.95	0.82
Recall (1)	0.90	0.76
F1-score (1)	0.92	0.79
Support (1)	4397.00	280.00
Accuracy	0.92	0.89
Macro Avg	0.92	0.87
Weighted Avg	0.92	0.89

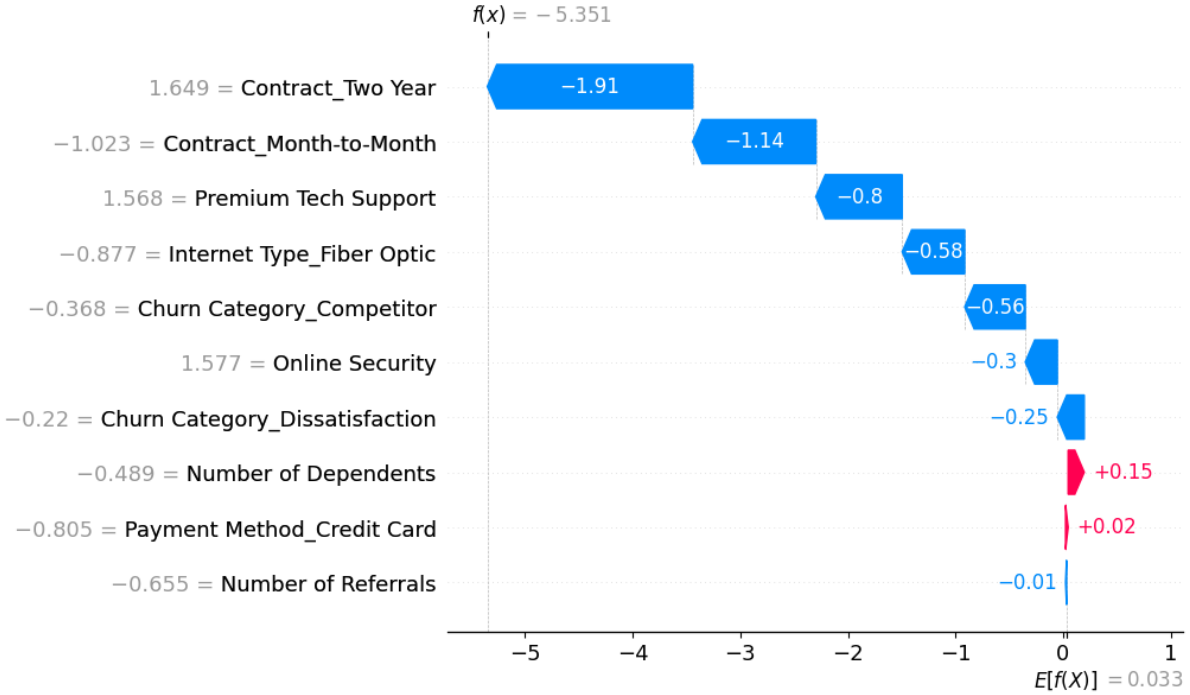
The model performance of XGBoost on train and test shows, the model had performed well on unseen test data. It scored 92% accuracy on the train set whereas it got reduced to 89% when predicting on test set, which is expected from a machine learning model. The more important observation is that, the model is not overfitting and generalizing well on new data. It was able to predict churn with 95% precision and F1 score of 92% on train set, while it predicted churn



on test set maintain precision of 82% and f1 score of 79%. The ROC curves further illustrate the model accuracy on train accuracy and test set over the training time. Which also indicates model's generalization on unseen data. Hence, with such balanced and consistent performance XGBoost has been considered as the most appropriate model for predicting churn. However, the scope of improvement still lies in the model and can be further optimized with large data sets.

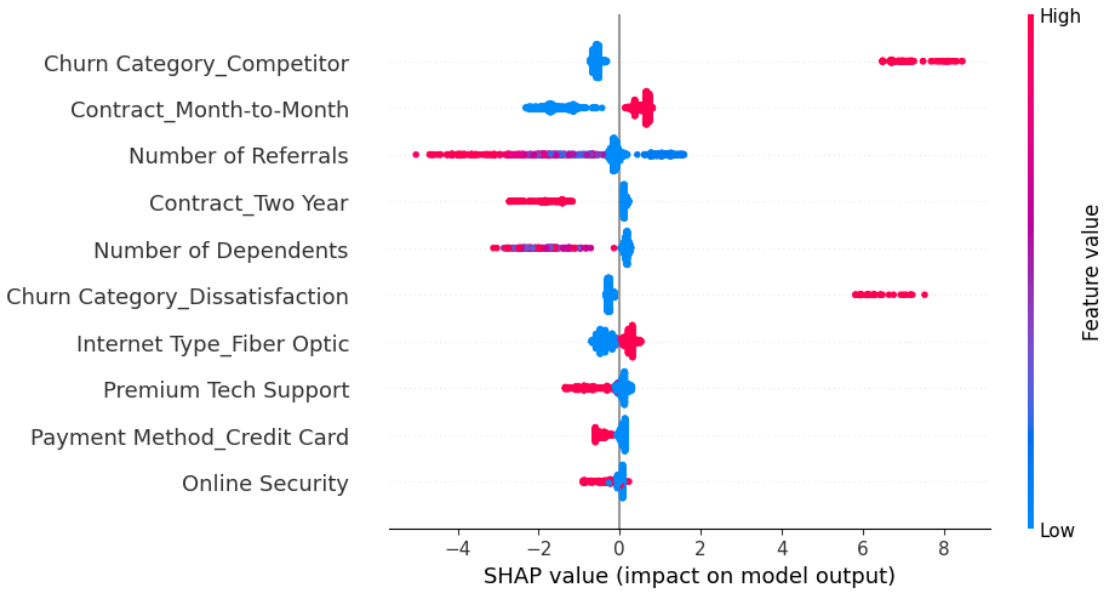
CHAPTER 4
DISCUSSION

The main objective of the study was not to predict customer churn only. Its main focus is to identify the key factors that impacts on customer to churn. These factors should be considered by the telecommunication company while developing and implementing business strategies. Therefore, explainable AI (XAI) model SHAP (Shapley additive explanation) was implemented on the XGBoost model for the purpose of interpreting which features drives customers churning decision.



Here, the SHAP summary plot shows impact of each feature on churn prediction. A negative value indicates the features reducing the predictions whereas a positive value represents a feature increasing the prediction accuracy. Based on this, it is visible that two-year contract significantly lowers the prediction score which describes that customers with a two-year contract tends to churn less. Additionally, month to month contract, premium tech support, fiber optic internet type, online security availability are the key factors that reduces customer to churn. So, telecommunication companies should focus on these factors and optimize their strategies and execution ensuring the best quality of service in these criteria. Moreover, people having a greater number of dependents are more likely to churn. This shows the company's lack of providing family packages; therefore, people switch to other service providers for obtaining their whole need. Another feature that increases customer churn is credit card

payment method users, the reason might be issues while paying bill or limited payment method used by the company. The company should concentrate on overcoming these pitfalls.



The bees warm plot further illustrates the feature importance and their SHAP value. The plot is indicating customers generally churn due to better service of competitors and dissatisfaction of the current products and services, indicated by red dots. Customers having Fiber optic internet connection with month-to-month contract and higher number of referrals tend to have lower SHAP values (blue), suggesting a lower likelihood to churn. Hence, the model's predictive accuracy is backed up by practically logical features that actually can impact a customer's decision on continuing a telecommunication service. Implementing the model on a large data might uncover many influential factors that can significantly improve customer relationship management strategies.

CHAPTER 5
CONCLUSION

In conclusion, this study has successfully developed a decent model for predicting customer churn and identified critical factors influencing customer decisions to leave. The telecommunications company can enhance its customer retention strategies, ultimately leading to improved customer satisfaction and business performance. The combination of advanced feature selection, model optimization, and interpretability techniques has provided a comprehensive approach to understanding and addressing customer churn, setting the stage for ongoing improvements and success in customer relationship management.

The insights derived from the SHAP analysis can guide the telecommunications company in refining its customer retention strategies. Focusing on enhancing service quality in areas identified as critical for retention and addressing pain points associated with high churn rates can lead to improved customer satisfaction and loyalty.

REFERENCES

[1] A. S. R. A. M. Q. A. K. A. R. S. A. Qureshi, "Telecommunication subscribers' churn prediction model using machine," in *Digital Information Management (ICDIM), Eighth International Conference on IEEE*, 2013.

[2] C.-H. J. J. L. K. Kim, "Improved churn prediction in telecommunication industry by analyzing a large network," *Expert Systems with Applications.*.

[3] L. H. W. C. H. K. C. Kirui, "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining," *International Journal of Computer Science Issues (IJCSI).*