MORINGA
Discover · Grow · Transform

MORINGA SCHOOL

Predicting Customer Churn in SyriaTel (Telco): A Machine Learning Approach

Presented By Kelvin Njenga

TABLE OF CONTENTS

INTRODUCTION

Customer churn in the Telecommunication industry is one of the major challenges, as it not only results in lost revenue but also increased marketing and customer acquisition costs. Churn can be defined as the loss of subscribers due to some underlying reasons.

To mitigate its impact, companies must understand the reasons why customers leave and implement strategies to retain them such as developing classification models that try to predict when a customer will churn based on these underlying reasons.

BUSINESS UNDERSTANDING

Problem Statement

SyriaTel, a company in the telecommunication industry, has approached us with a pressing challenge: high customer churn rates are leading to financial losses and decreased customer satisfaction.

Main Objective

The goal of this project is to build a classification model that can accurately identify which customers are at risk of leaving the company(churning) and take proactive measures to retain them.

SPECIFIC OBJECTIVES

To identify the most significant features in determining whether a customer will churn or not

To build a logistic regression model that can accurately classify a customer churning or not based on input features from the Kaggle SyriaTel prediction dataset

To assess the performance of the various classification models and identify potential areas for improvement

METRIC FOR SUCCESS

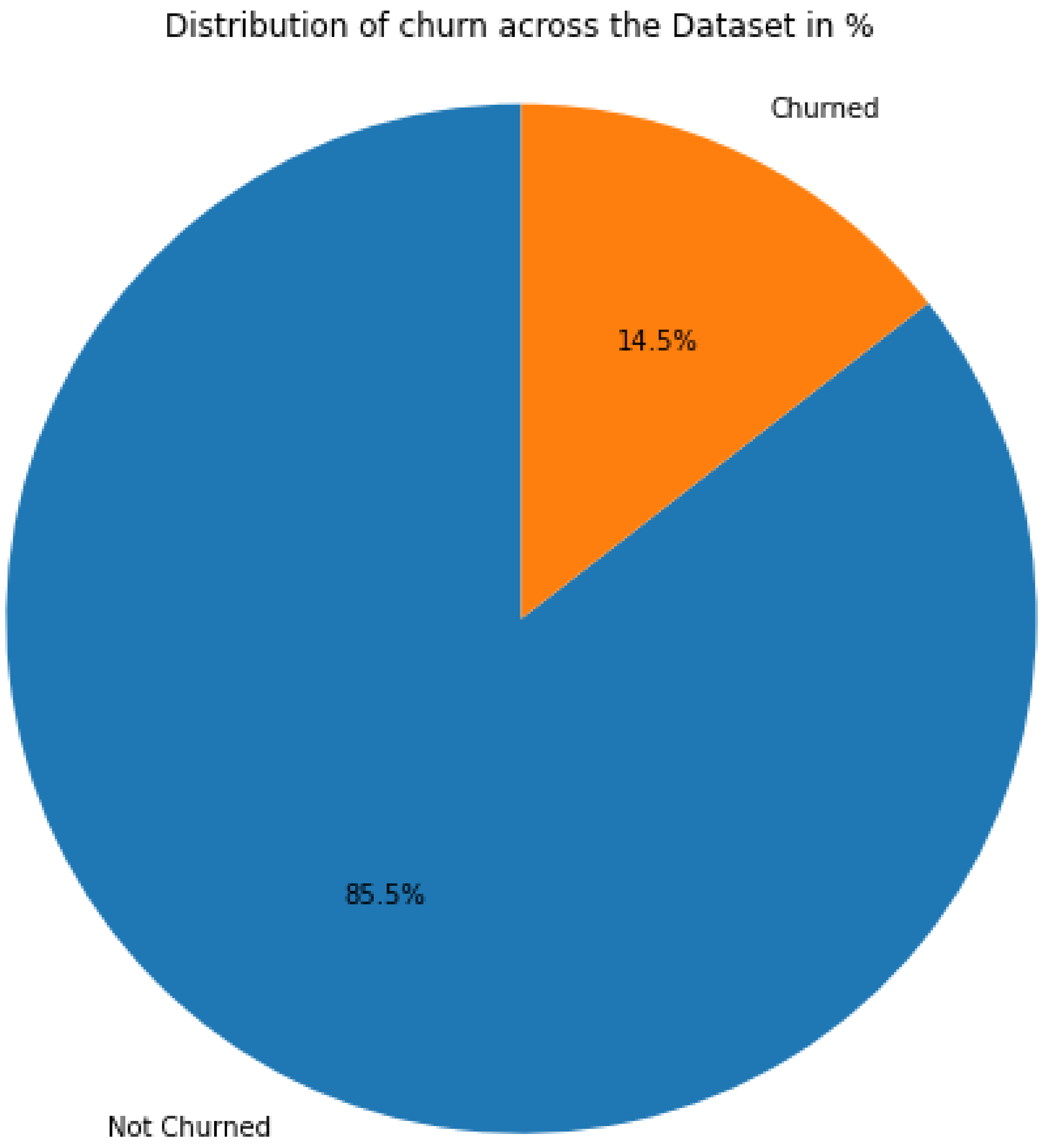
The project will be considered a success if the developed classification model is able to accurately identify a high proportion of actual churners(High Recall)

Demonstrate good generalization performance (Accuracy of 80% on unseen dataset).

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

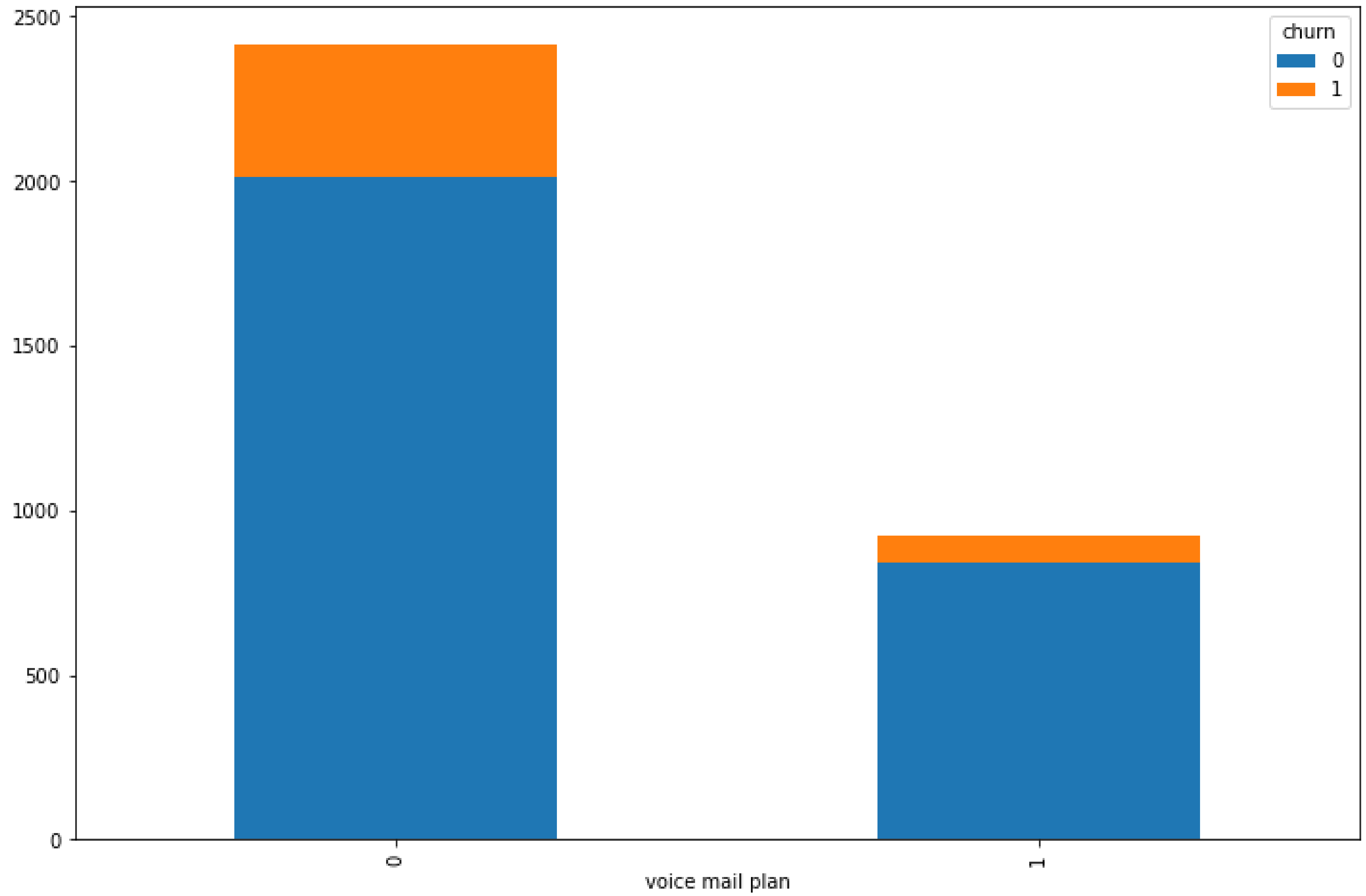
a. What is the distribution of customers that churn



Based on our pie chart 14.5% of the population has already churned while the majority (85.5%) have not yet churned.

BIVARIATE ANALYSIS

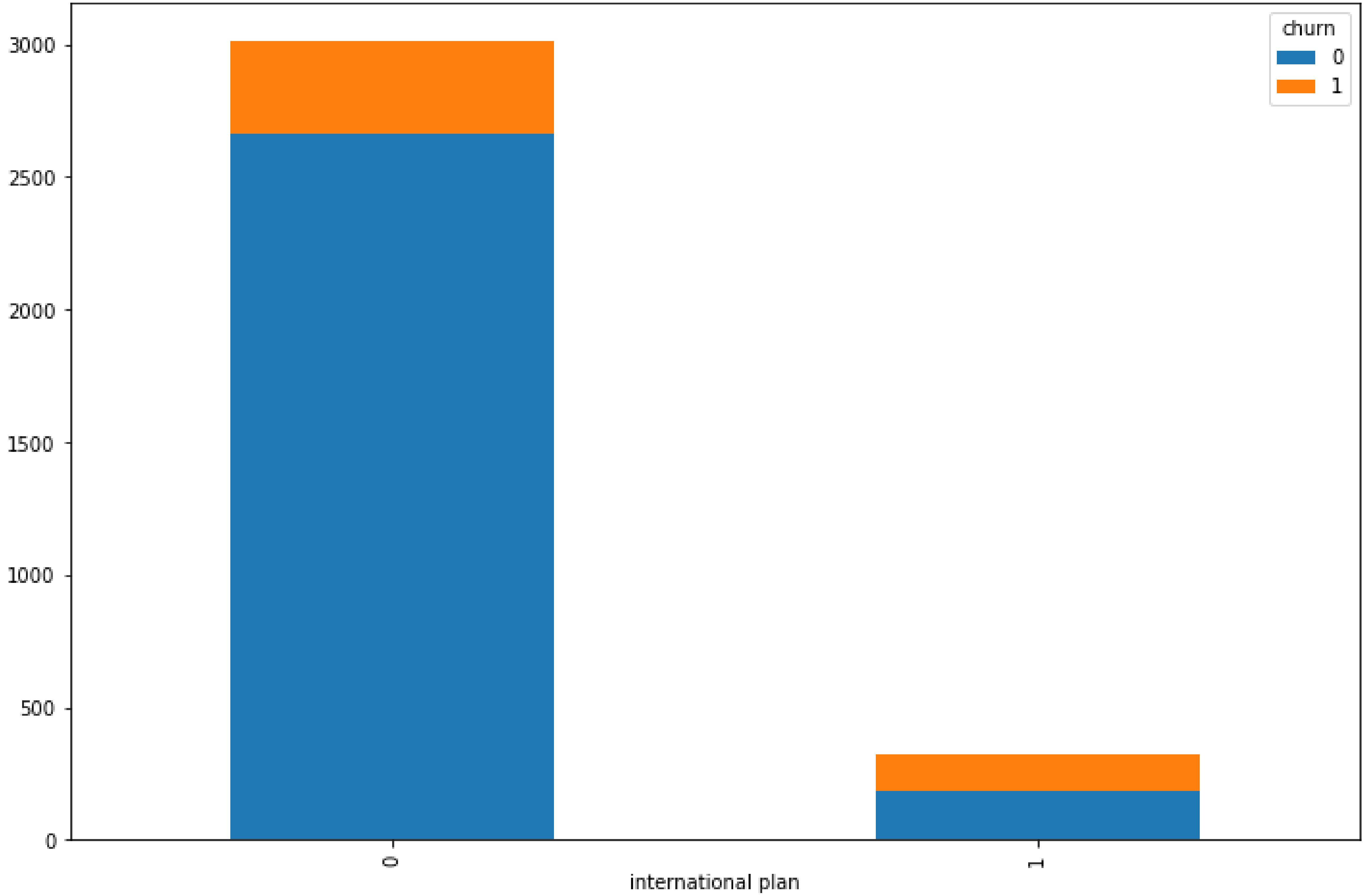
a. Are customers subscribed to a voice mail plan likely to churn?



Customers subscribed to the voice mail plan: 27.66%

Percentage of subscribed customers who churned with voice mail plan: 8.68%

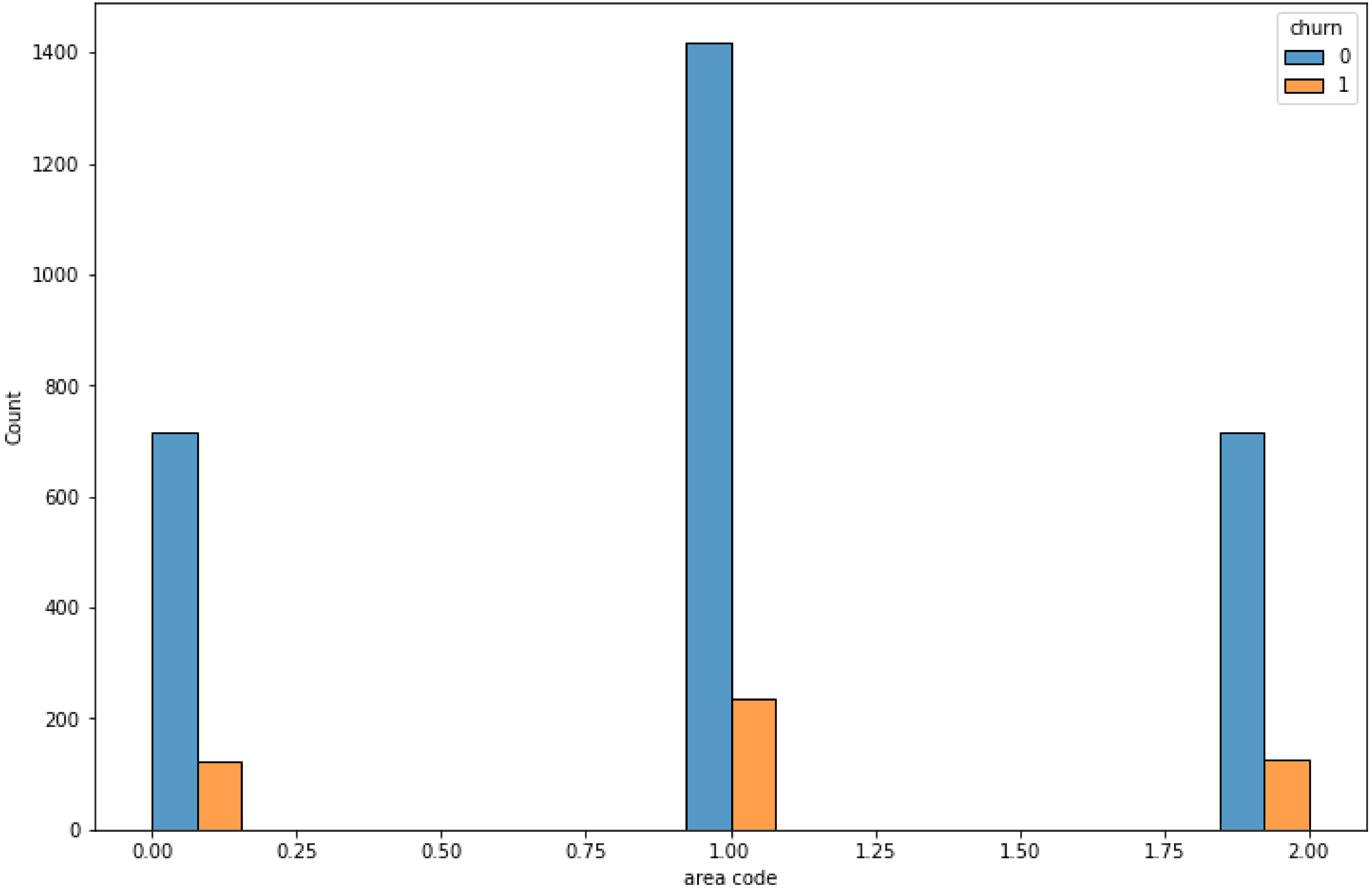
b.Are customers subscribed to a International plan likely to churn?



Customers subscribed to the international plan: 9.69%

Percentage of subscribed customers who churned with an international plan: 42.41%

c. What area code with the highest churn rate?



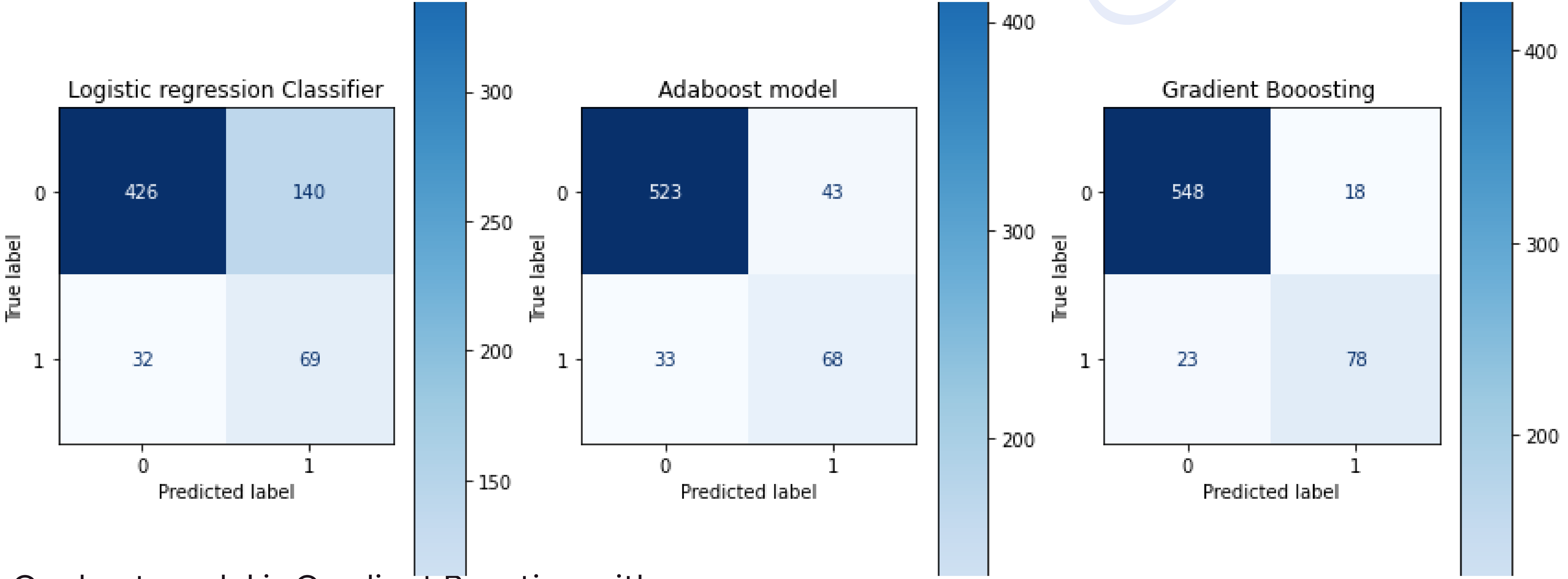
Area code 415 represented as 1 has the highest retention rate compared to the rest

The area code with the highest churn rate is: 510 represented by 2

DATA MODELING

MODELS PART 1

a. Model Diagnostics



Our best model is Gradient Boosting with :

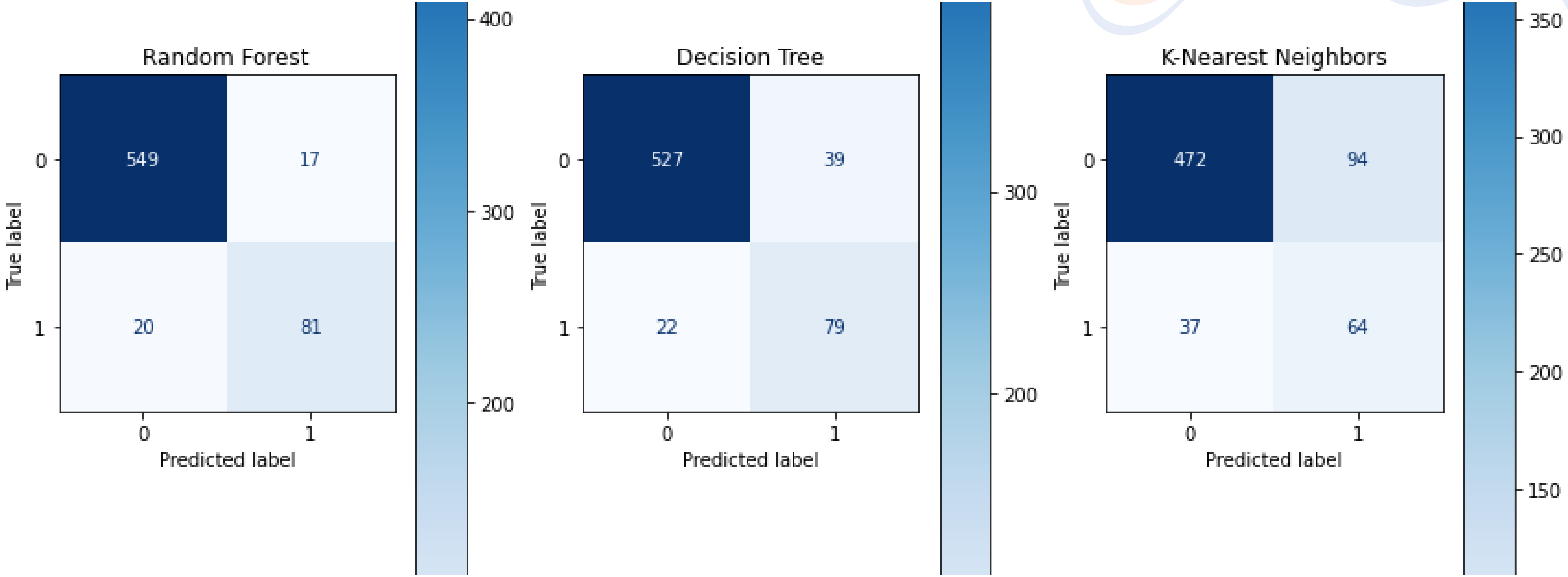
Accuracy: 0.9385307346326837

F1-Score: 0.7918781725888325

Recall: 0.7722772277227723

MODELS PART 2

a. Model Diagnostics



Our best model is Random Forest with :

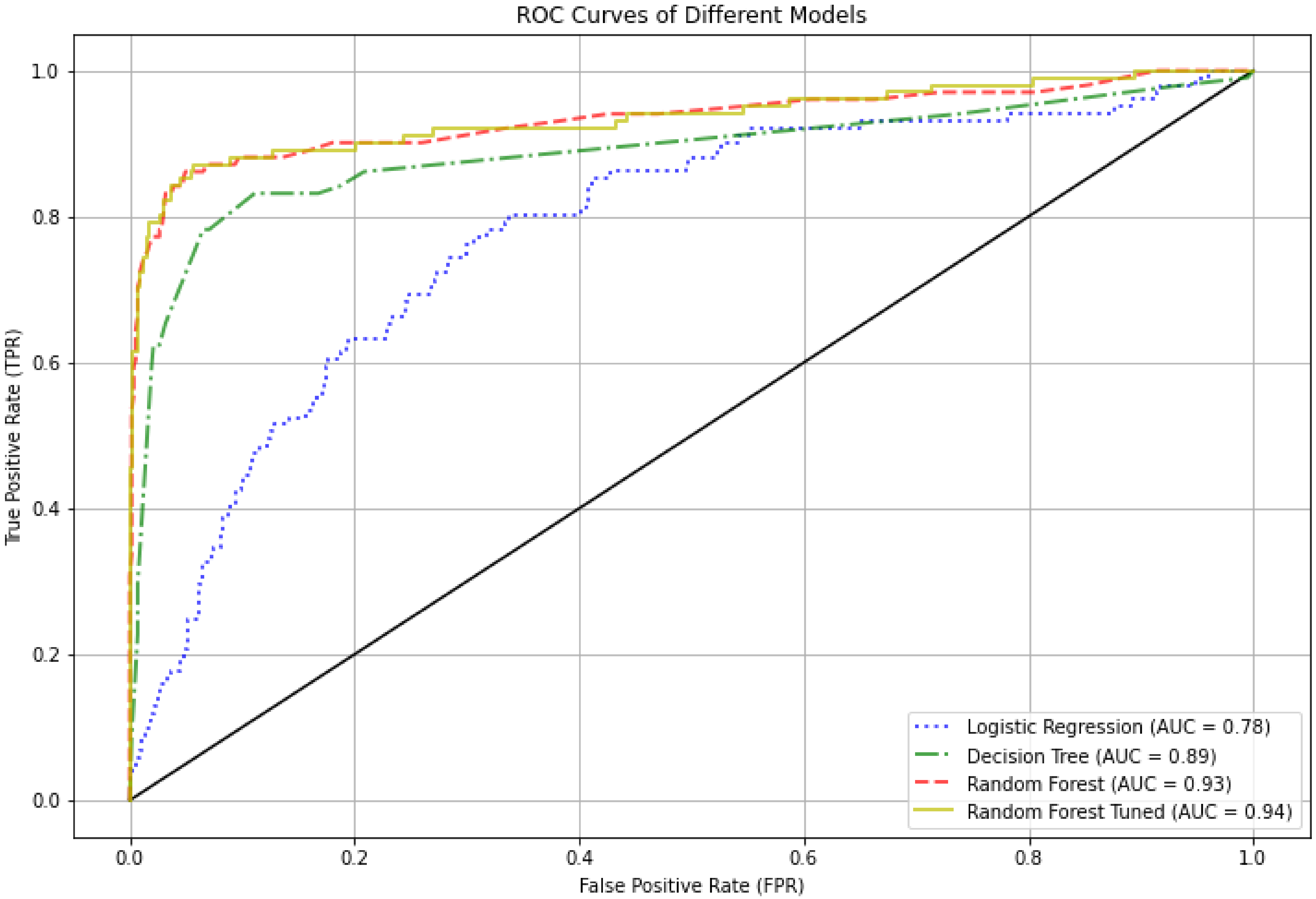
Accuracy: 0.9445277361319341

F1-Score: 0.814070351758794

Recall: 0.801980198019802

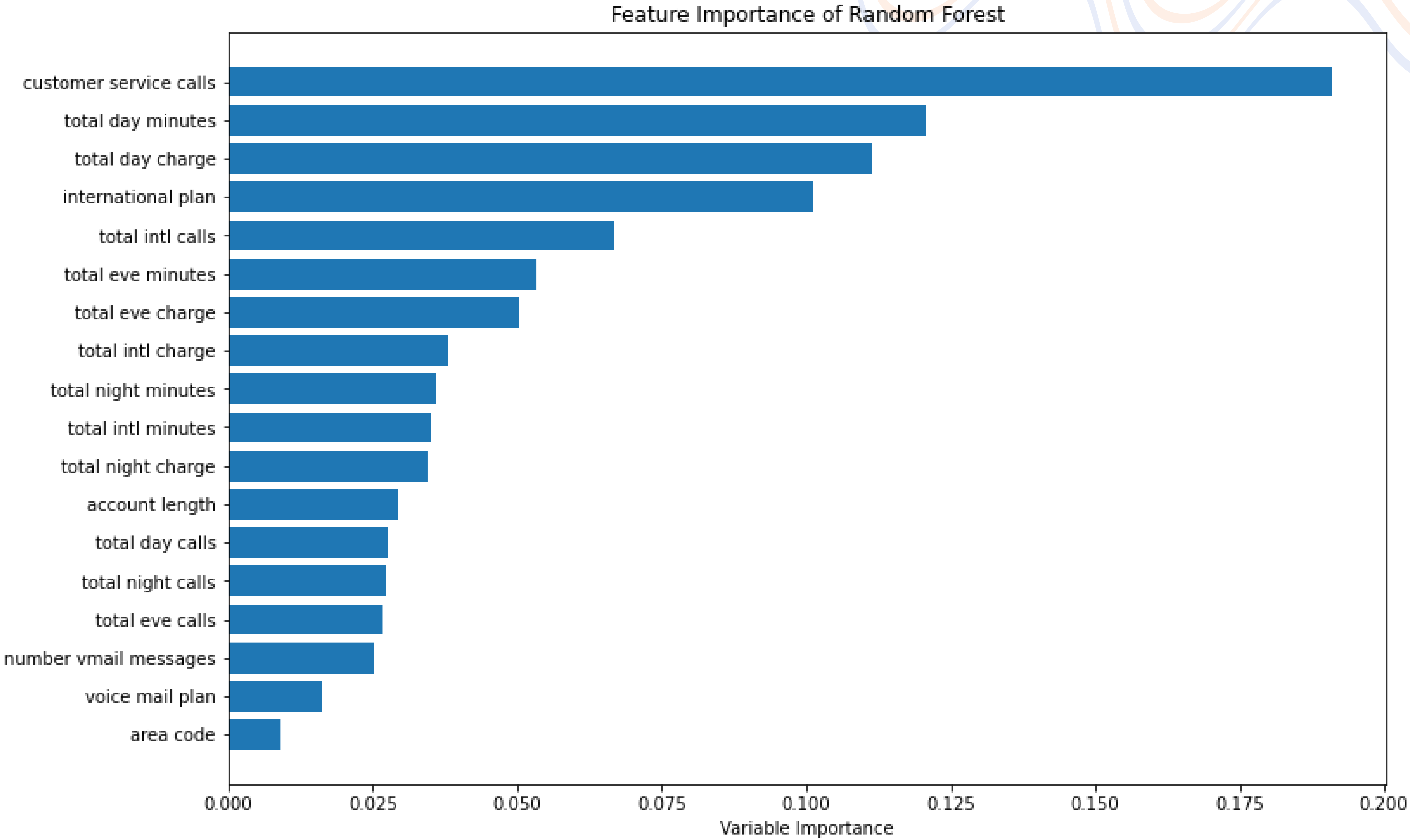
FINAL MODEL

Best Overall Model



Based on our ROC curve we can conclude that both the Random forest and tuned random forest models are our best models since they have an AUC of around 0.93 and they are hugging the top left side of our graph, they also have a high recall of around 0.82 when predicting a customer will churn as compared to the other models

FEATURE IMPORTANCE



SUMMARY

Logistic regression Model

The Baseline Model has an accuracy of 74.2% with a recall of 68.3%, its performance is relatively low compared to the metrics set however, it does not overfit or underfit.

Adaboost Model

The Adaboost model has an accuracy of 88.6% and a recall of 67.3% its recall is also low compared to the set metrics. This still wasn't a perfect fit.

Gradient Boosting model

The gradient boosting model has an accuracy of 93.1% and a recall of 77.2% its performance is high as it meets the set metrics. Therefore it's a good model to use in the classification of churn.

SUMMARY

Random Forest Model

The Random Forest Model has an accuracy of 94.5% with a recall of 80.1%, its performance is high as it meets the set metrics. Therefore this is the best model to use

Decision Tree Model

The Decision tree model has an accuracy of 90.9% and a recall of 78.2%. its performance is high as it meets the set metrics. Therefore it's a good model to use in the classification of churn.

K-NearestNeighbor model

The KNN model has an accuracy of 80.3% and a recall of 63.4% its recall is also low compared to the set metrics. This still wasn't a perfect fit.

CONCLUSION

In conclusion, the results of the mean random forest cross-validation score (k=5) on predicting the churn rate showed an accuracy of 0.95, with a weighted average of 0.95. The precision, recall, and f1-score for class 0 (not churned) were 0.97, 0.97, and 0.97, respectively. For class 1 (churned), the precision, recall, and f1-score were 0.82, 0.82, and 0.82, respectively. The macro average and weighted average for precision, recall, and f1-score were 0.89 and 0.95, respectively.

These results indicate that the random forest model performed well in terms of accuracy, with a high weighted average for precision, recall, and f1-score. The model's performance in predicting class 0 (not churned) was slightly better compared to class 1 (churned), with a higher precision, recall, and f1-score. Overall, the random forest model was considered the best model for predicting the churn rate.

RECOMMENDATIONS & FUTURE IMPROVEMENT IDEAS

An insurance company should consider:

1 Enhance the prediction of churn by creating new features from the existing data. For example, a feature for the average daily or monthly charges

2 Offering loyalty programs: Offering incentives and rewards for customers who stay with the provider for a longer period of time can reduce churn

3 Offering competitive pricing and packages: Customers are more likely to switch to another provider if they feel that they are not getting good value for their money.

4 Collecting more data on other factors such as age, gender, Network Coverage and Quality to improve the accuracy of their churn model.

ANY QUESTIONS?

THANK YOU!