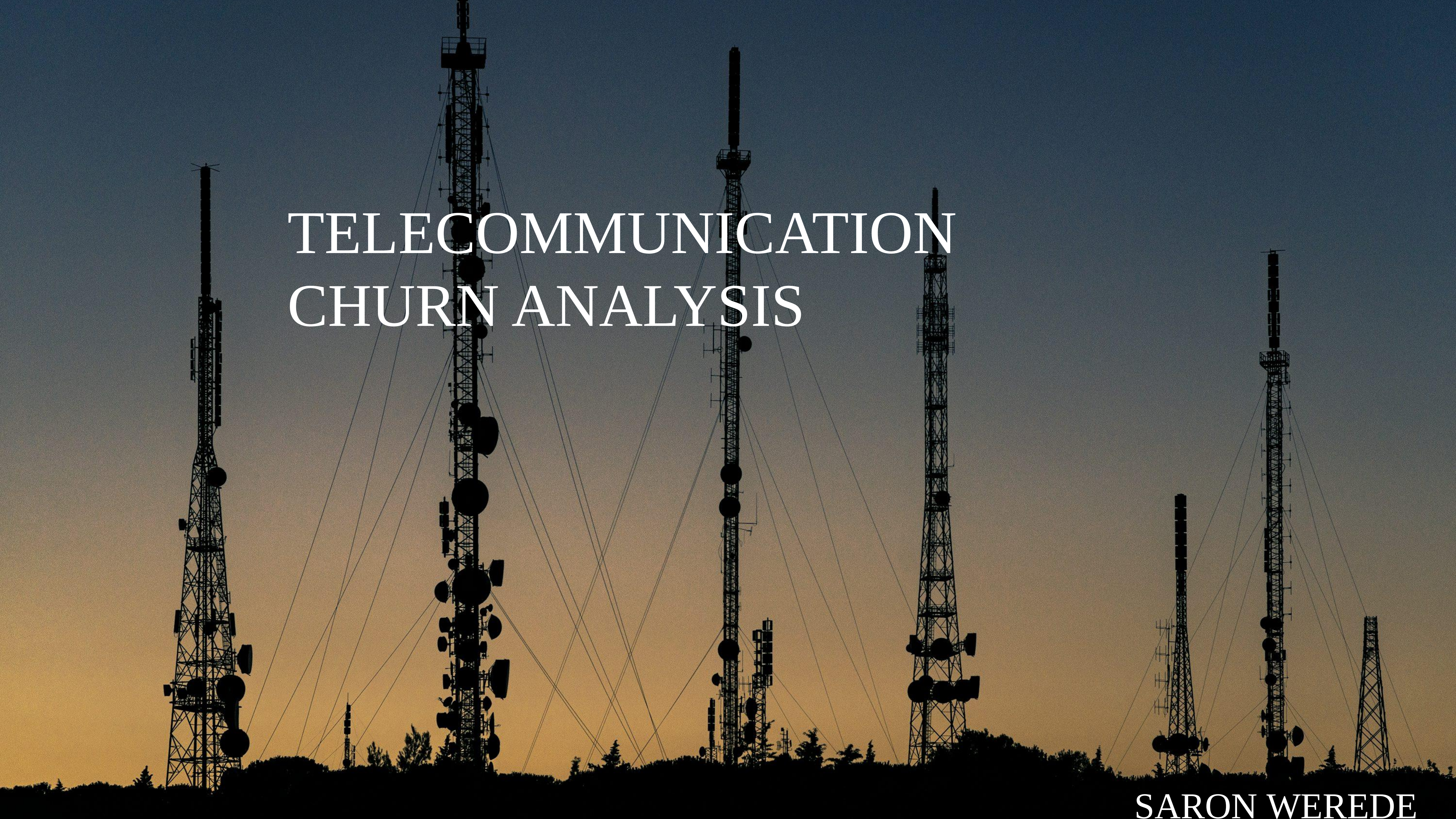


TELECOMMUNICATION CHURN ANALYSIS

SARON WEREDE

OVERVIEW

This presentation is divided in different parts:

- Business understanding
- Data understanding
- Modeling
- Conclusions

Business Understanding

Telecommunication market is expanding day by day. Companies are facing a severe loss of revenue due to increasing competition hence the loss of customers. They are trying to find the reasons of losing customers by measuring customer loyalty to regain the lost customers. The customers leaving the current company and moving to another telecom company are called churn

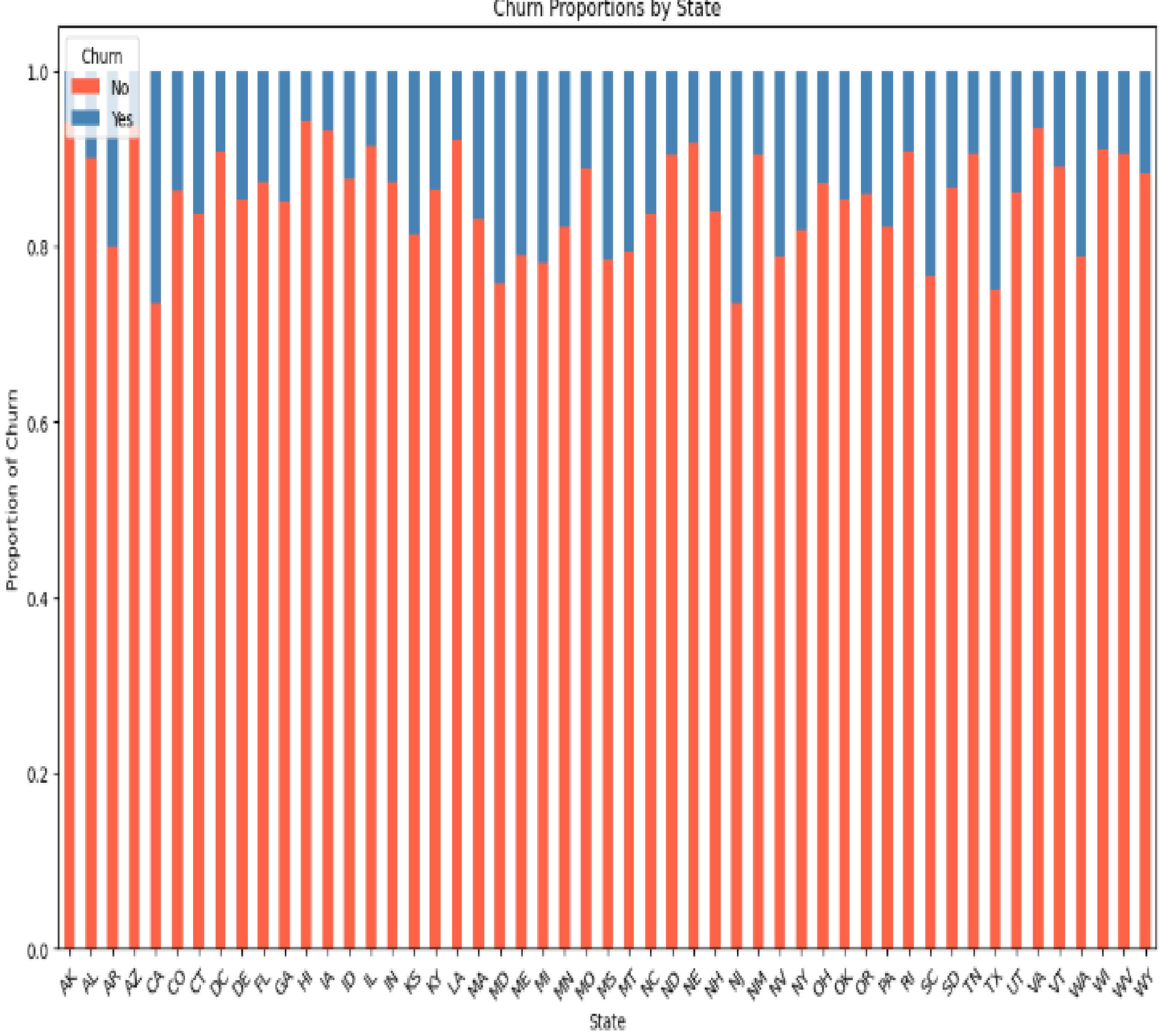
OBJECTIVE

- How to reduce churn and increase retention by predicting high-risk customers and what makes them churn?
- How can I optimize model performance with the right balance between precision and recall?
- Building models that Minimize false positives
- How class imbalance can be handled to ensure accurate predictions for churners?
- Develop targeted campaigns based on the model's predictions.

METHODS

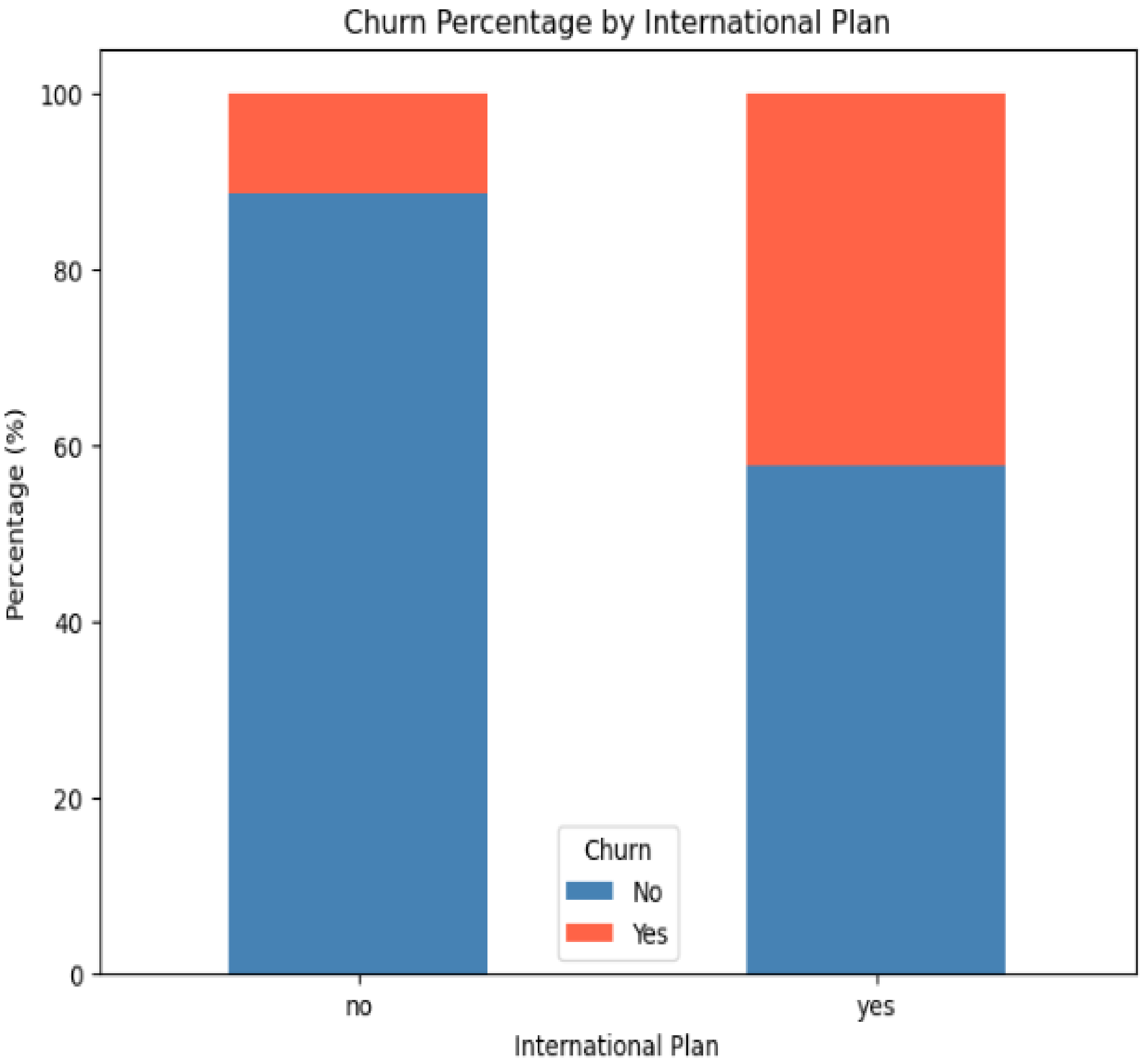
- Data collection
- Data cleaning
- Exploratory Data Analysis
- Data Visualization
- Data Modeling: logistic regression, SVM, KNN, random forest, CatBoots
- Model evaluation
- Understanding of selected features by the model

Data Understanding



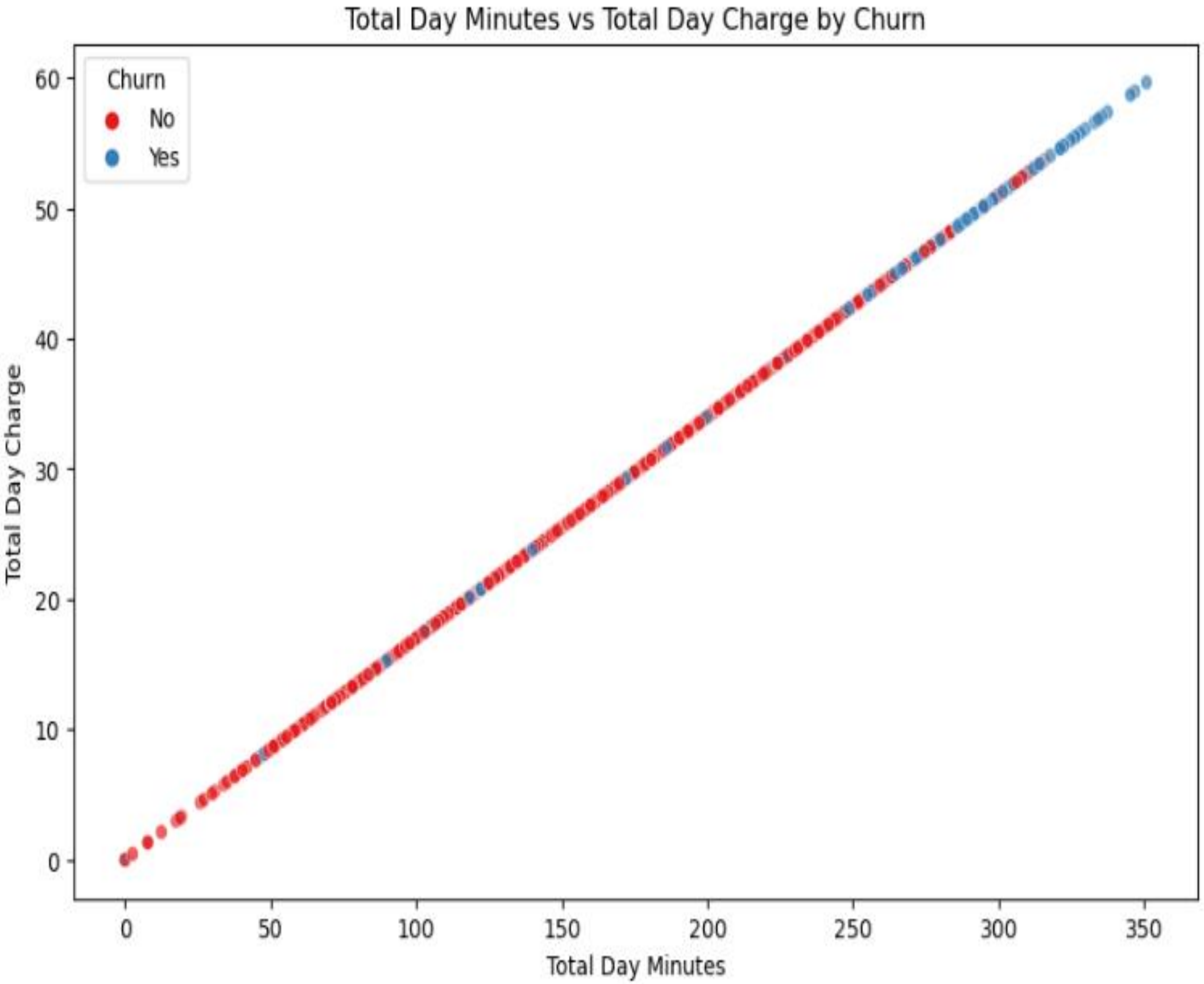
In this graph we can see the churn distribution by state in terms of percentage.

Conclusion: some states like NJ and CA have the highest churn



This plot explores the relationship between international calls and churn.

Conclusion: people that have international calls tend to churn more than the people who don`t have.

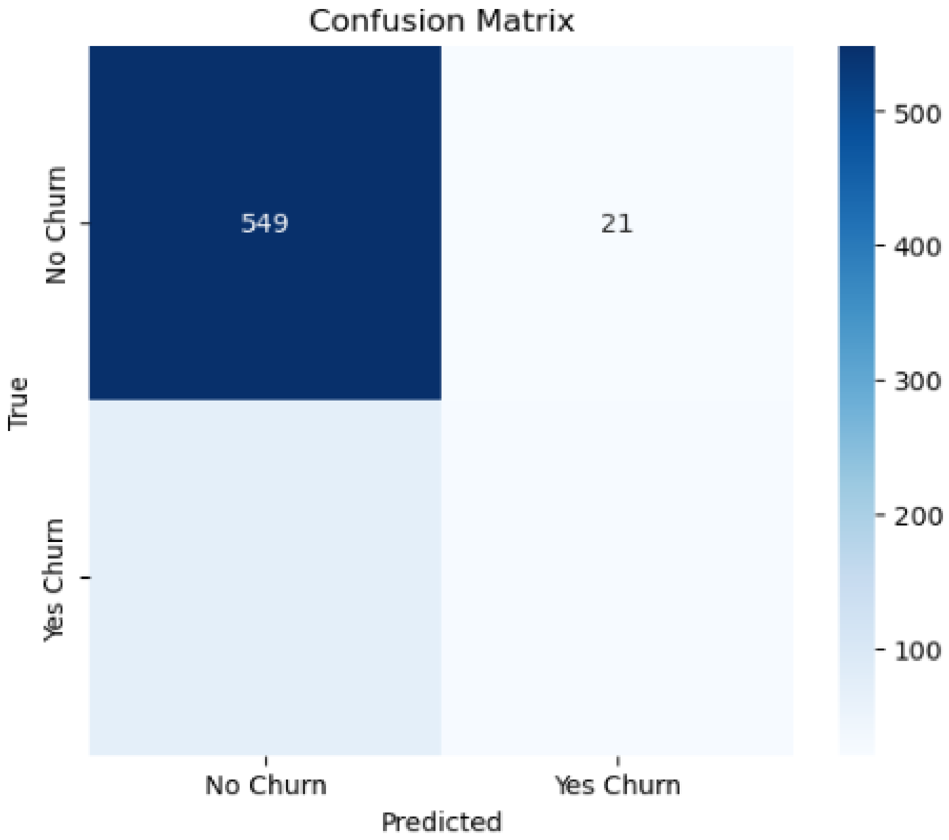


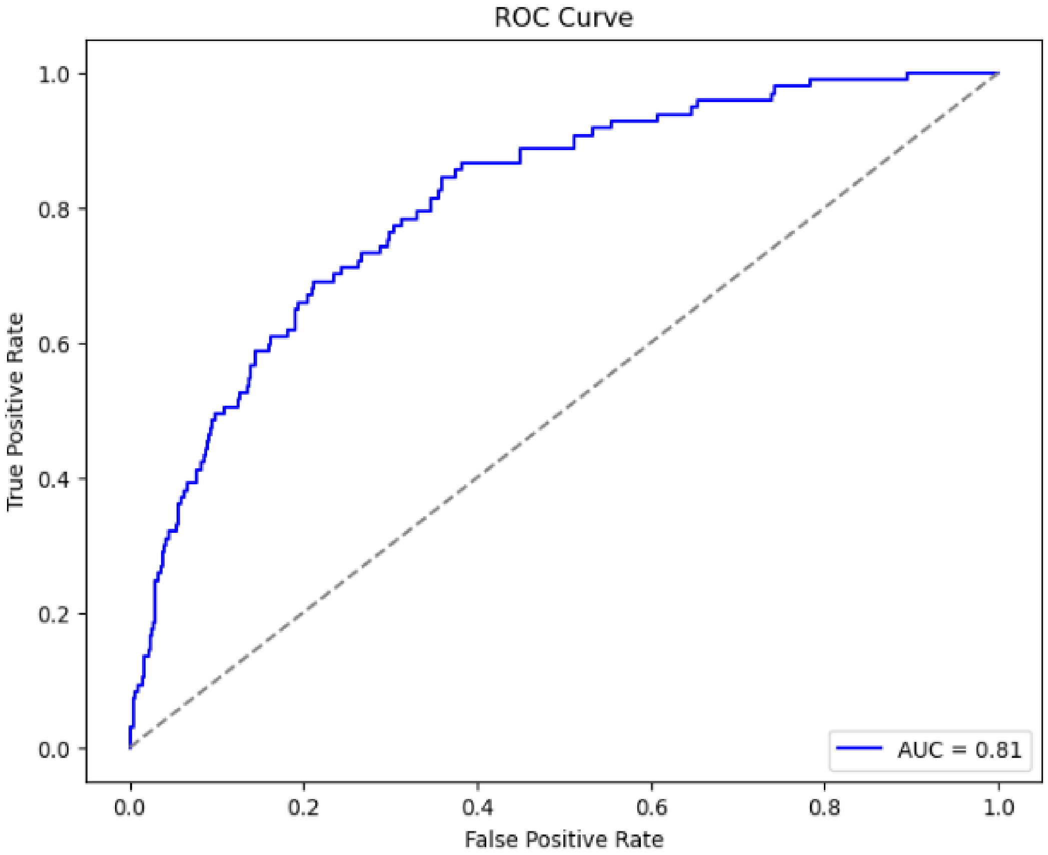
This plot shows,total day minutes feature against day charge feature, colored by churn.

Conclusion: the relationship is 1 to 1, and as the charges or minutes are high the people tend to churn more.

Data Modeling

Logistic Regression

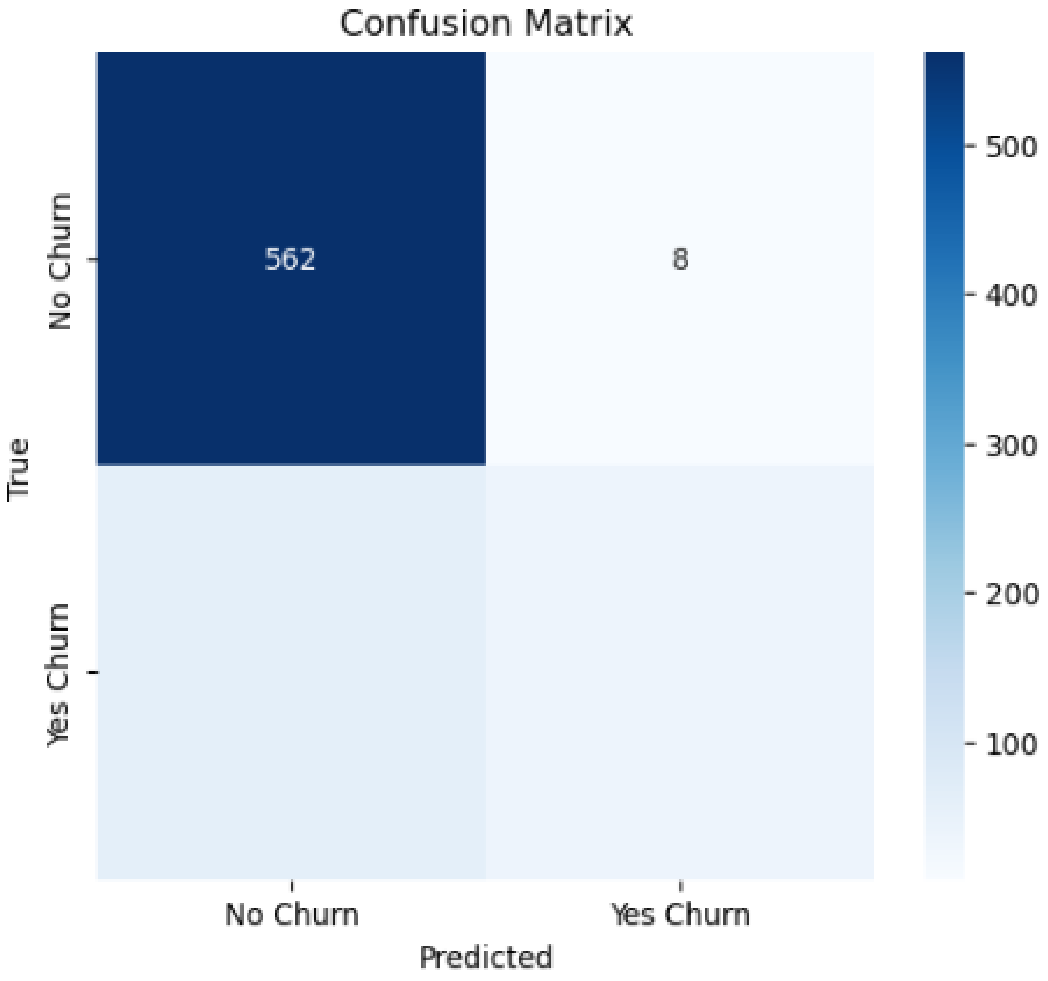


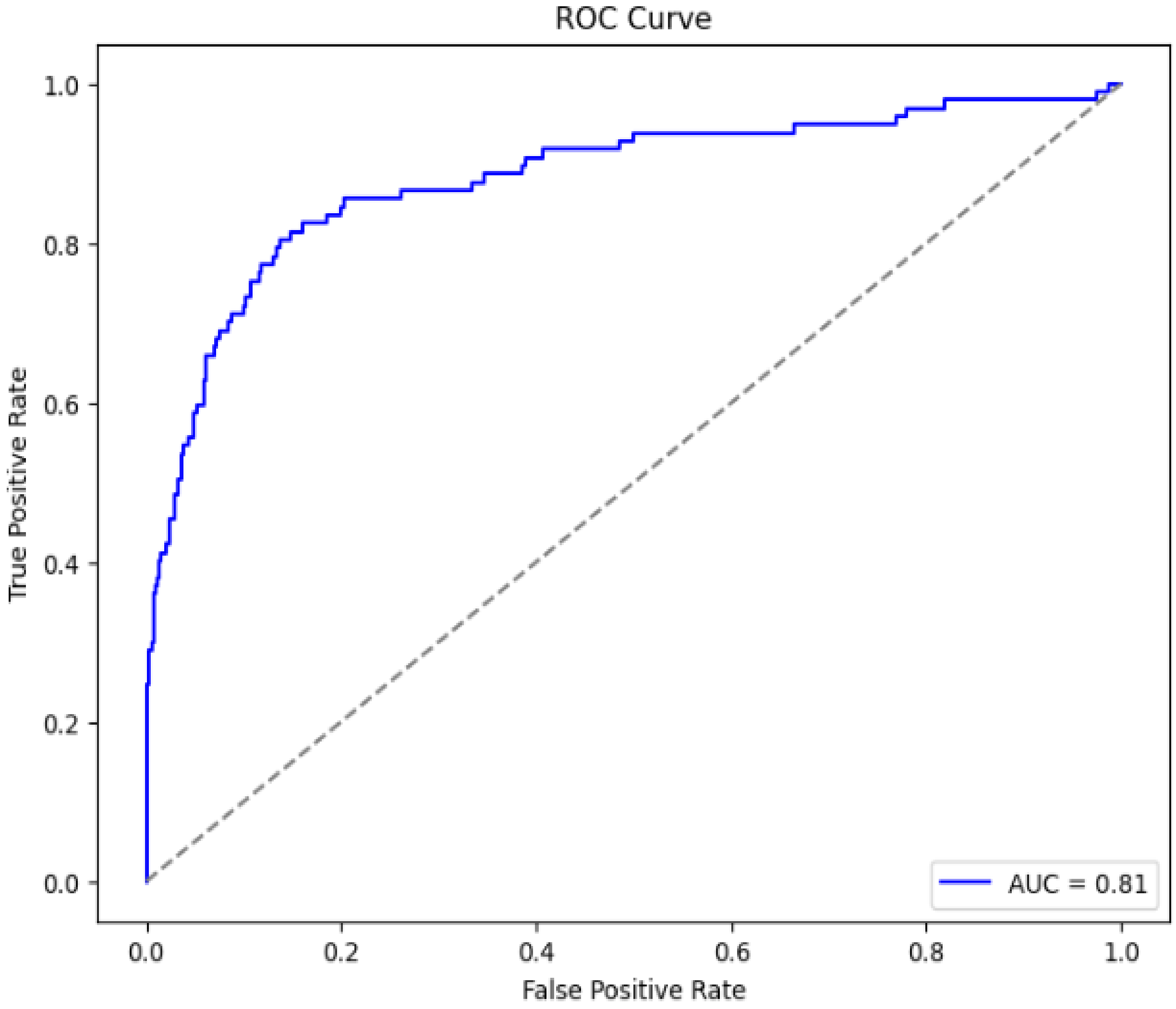


Conclusion:

- the model has an accuracy of 86% and a recall of 99% for no churn and a 26% for yes churn
- from the coefficients analysis ,total intl minutes thas the highest positive impact on the prediction, total evening calls have almost close to zero effect and total intl calls have the highest negative impacts.

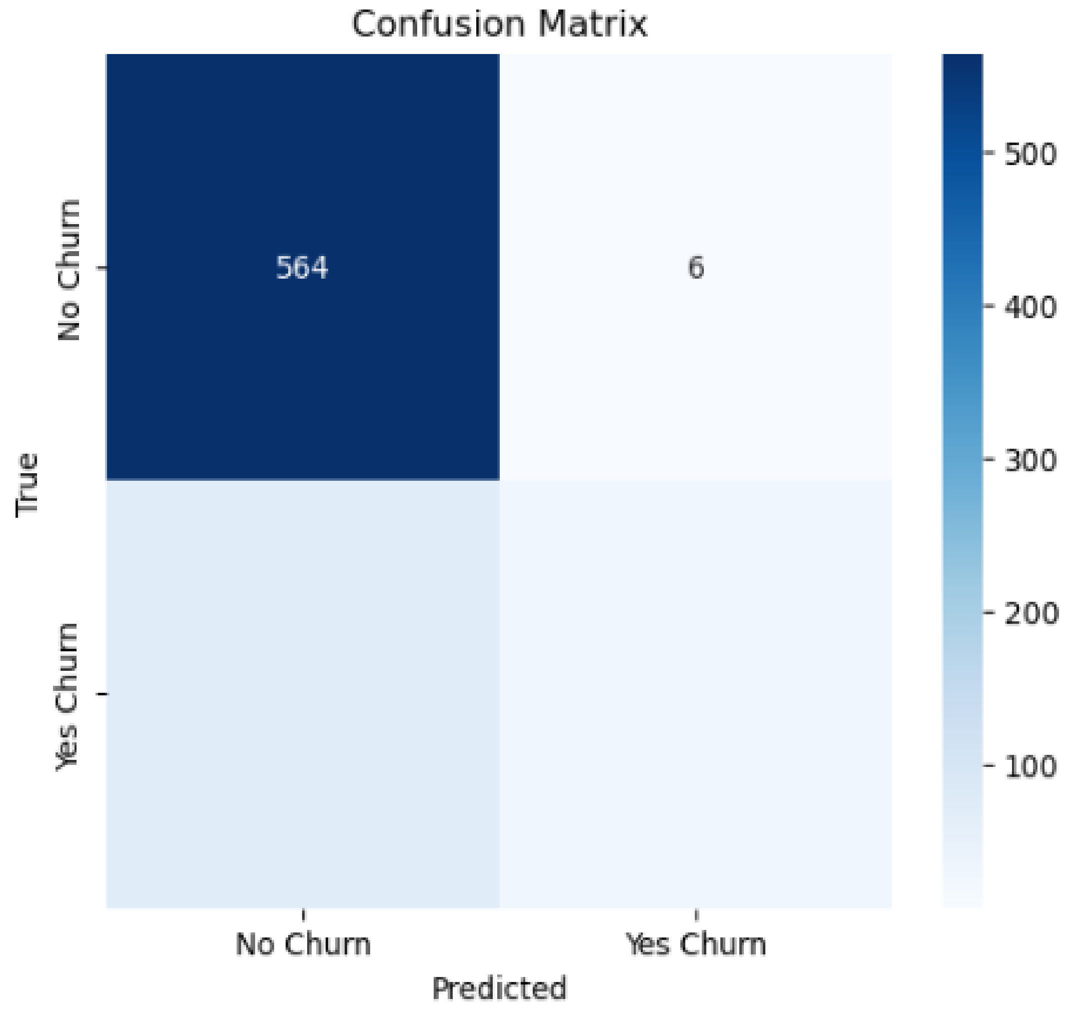
SVM Model

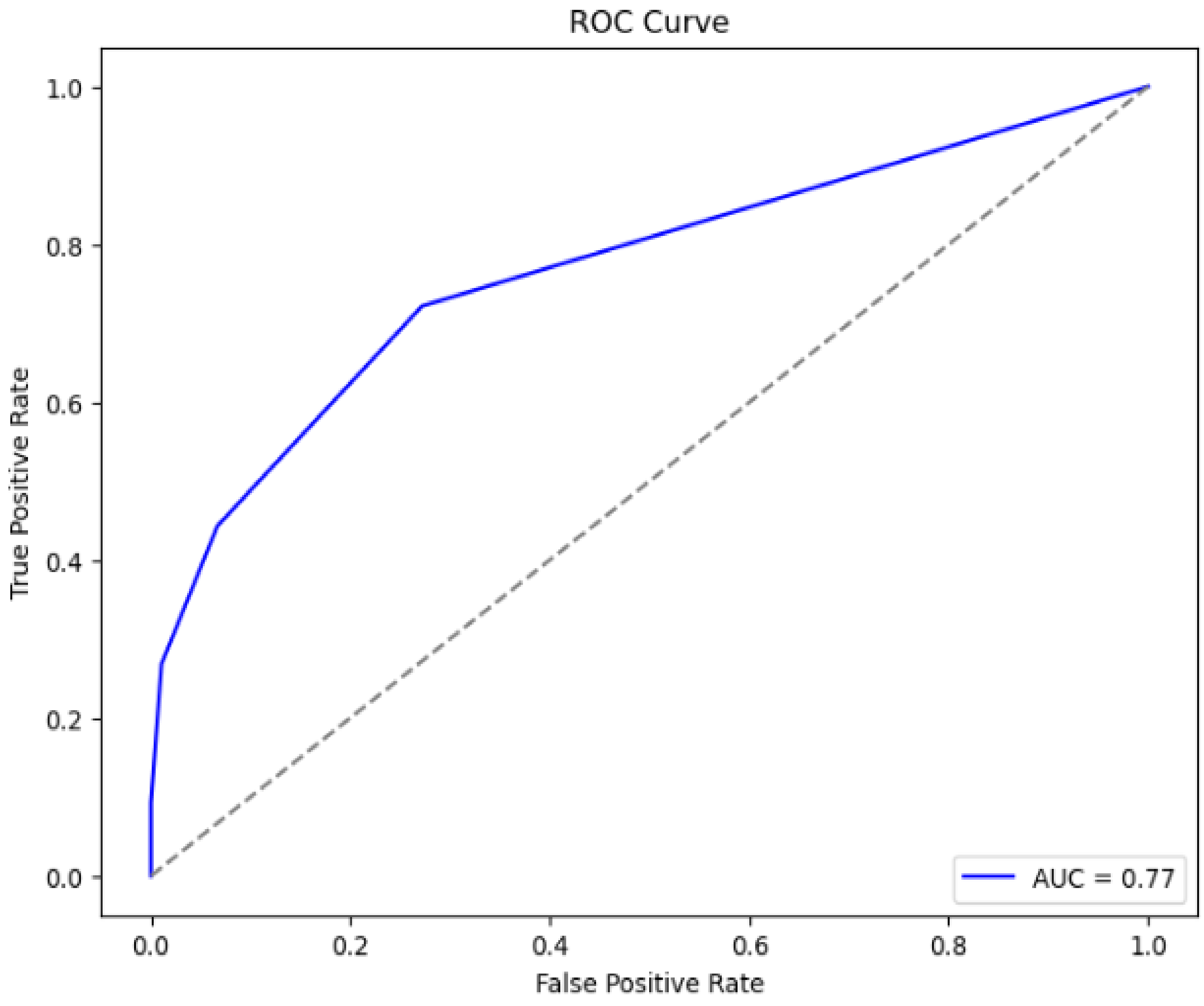




Conclusion: the accuracy is 90% and the recall of no church is 99% while for yes churn is 38%. We can observe that the model has improved in comparison to logistic regression.

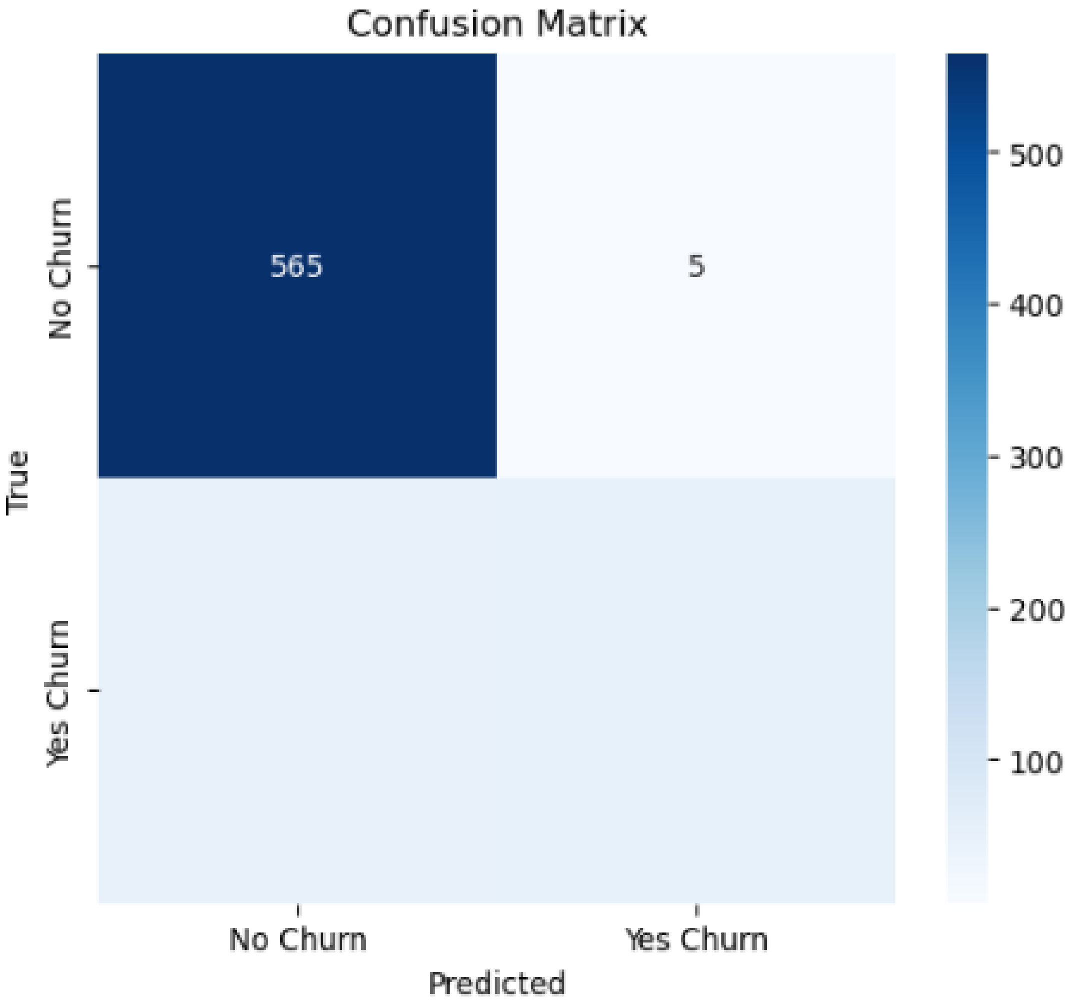
KNN Model

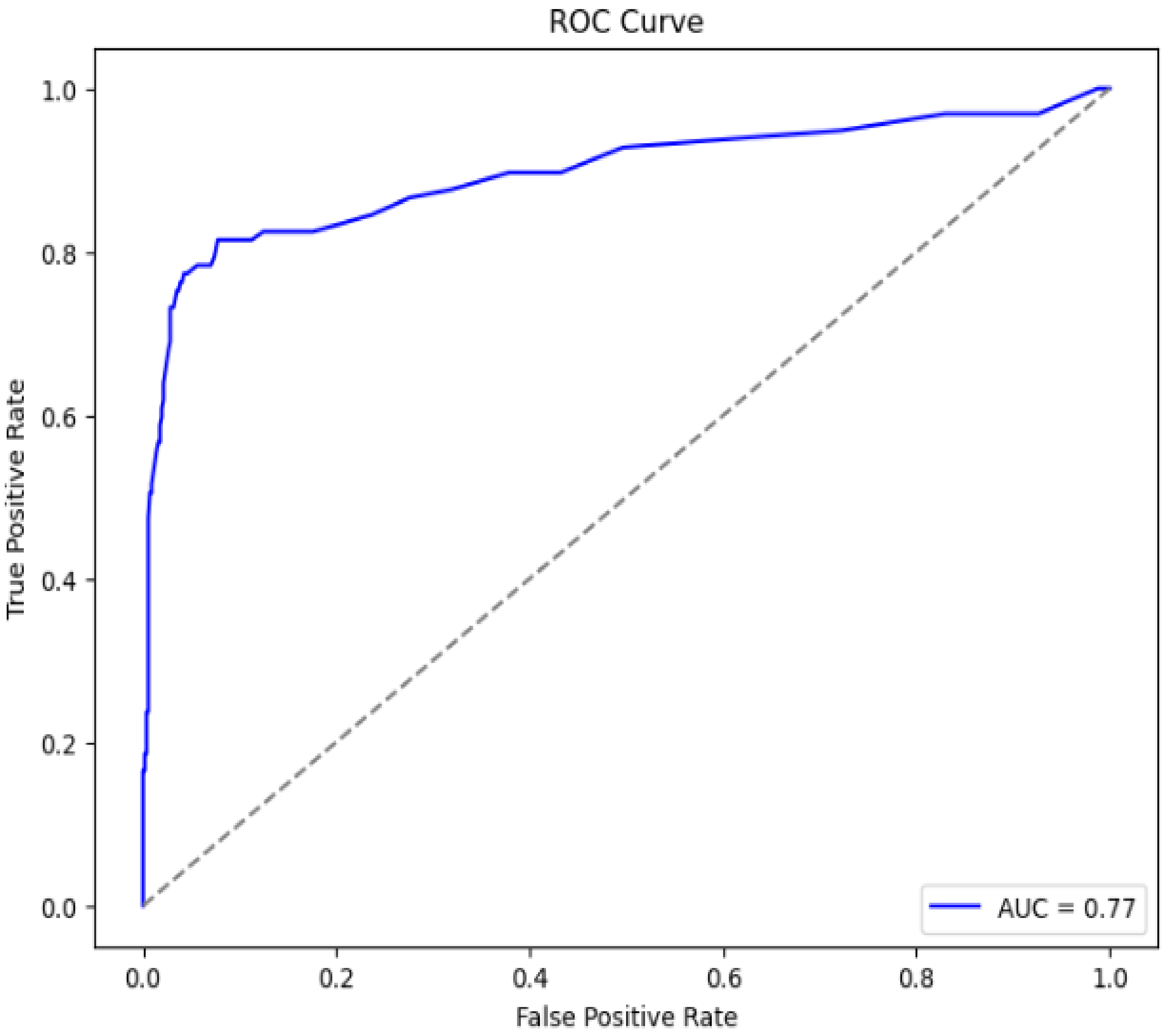




Conclusion:the accuracy is 88% and the recall of no church is 99% while for yes churn is 27%
SVM performe much better than KNN.

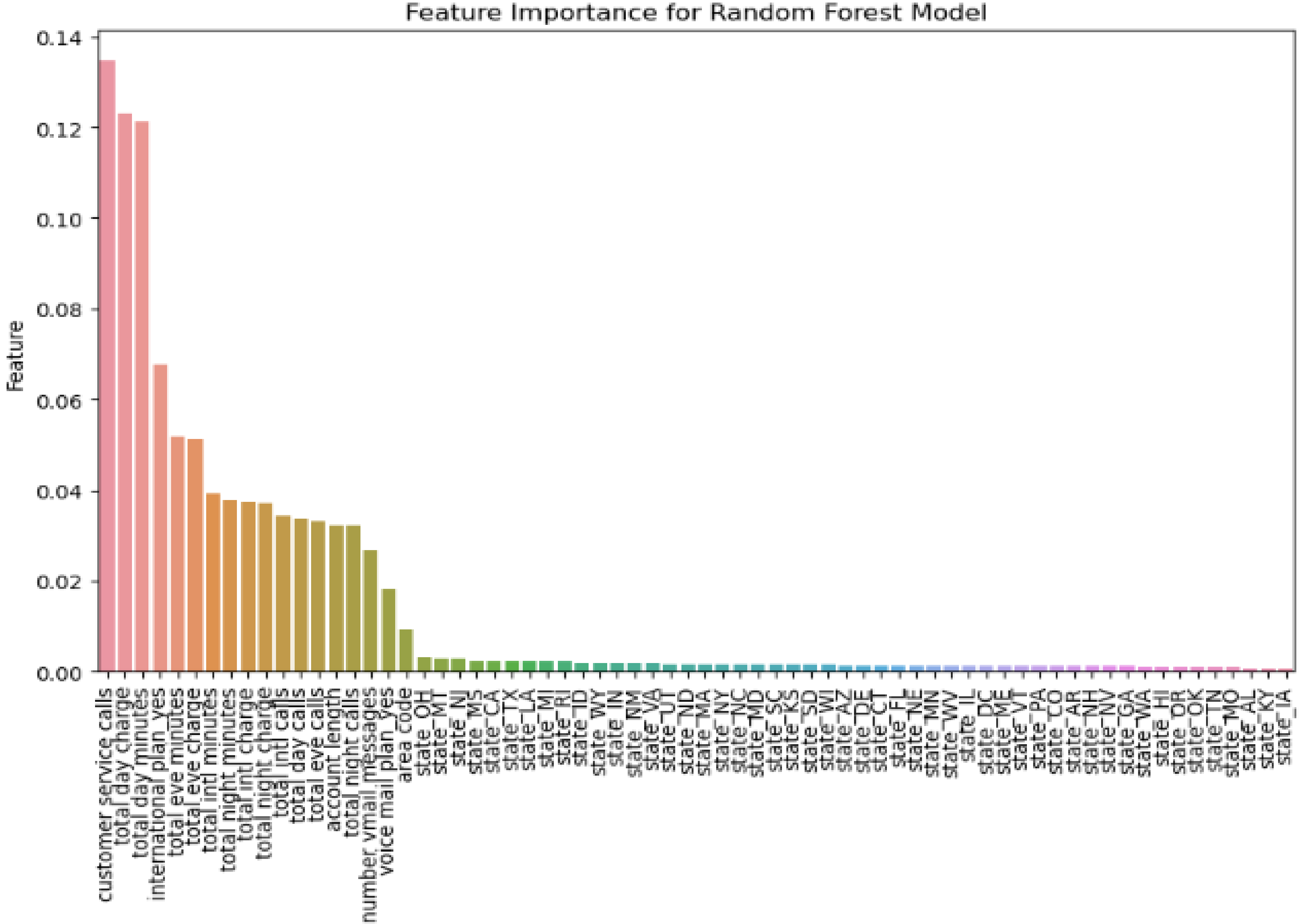
Random Forest





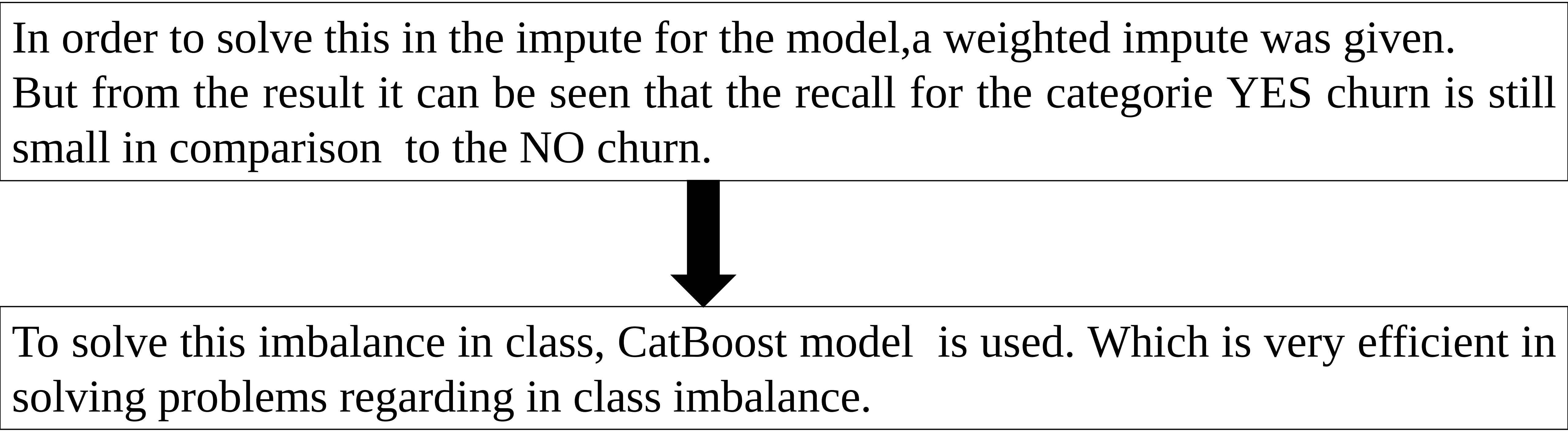
Conclusion:the accuracy is 92% and the recall of no churn is 99% while for yes churn is 59% Random Forest performe better than SVM.

Features important of Random Forest Model



From the feature importance it can be seen that customer service calls, total day charge and total day minutes have the most importance in estimating the churn.

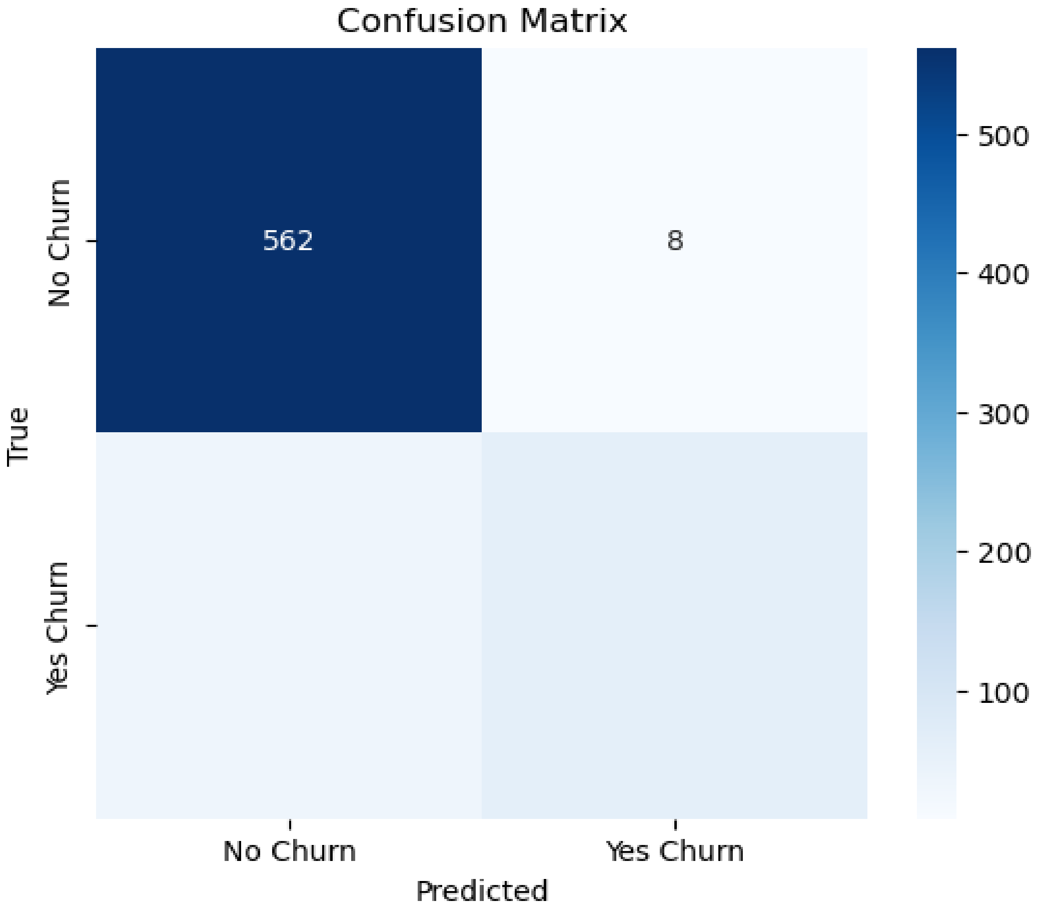
Issue with Modeling

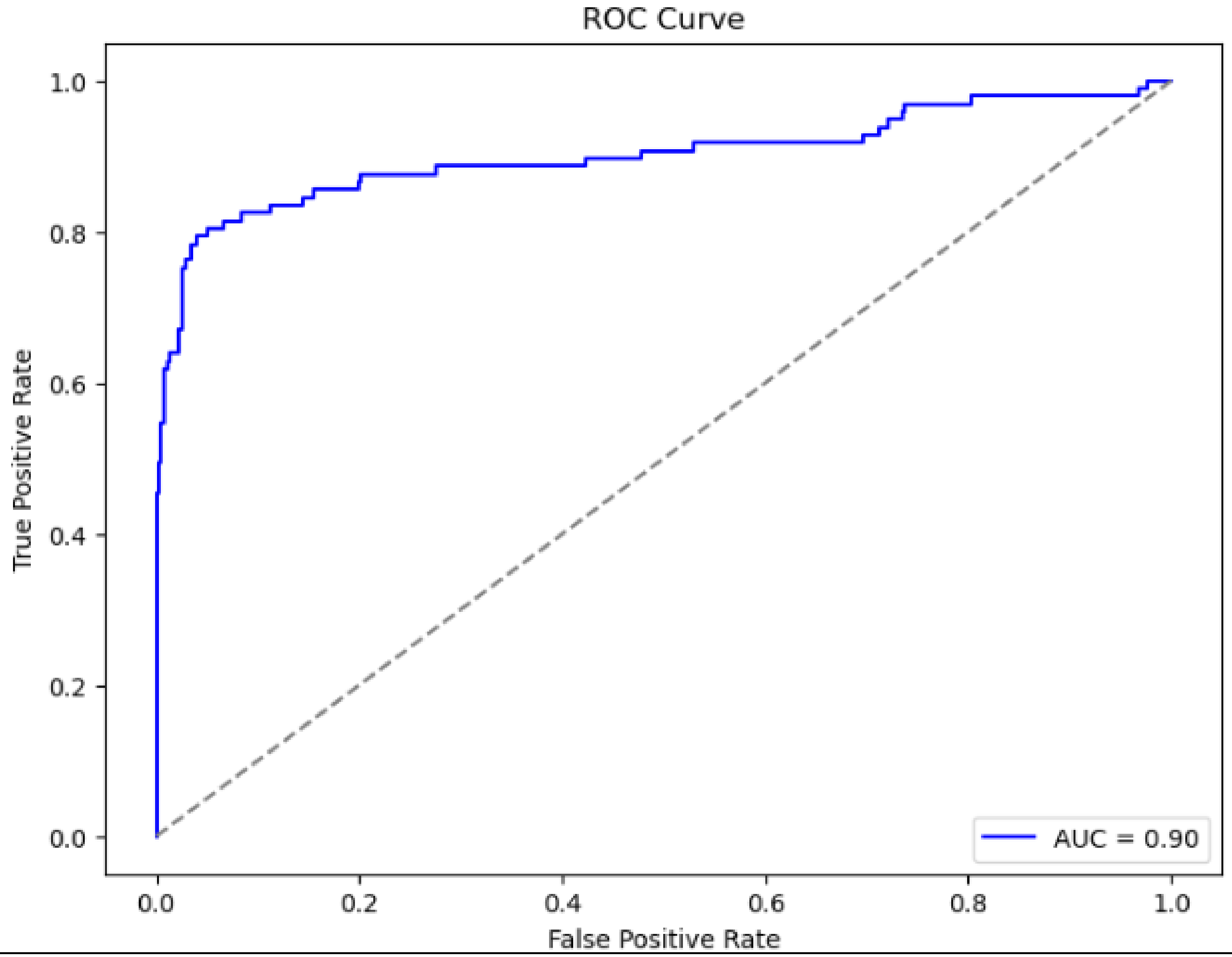


In order to solve this in the impute for the model,a weighted impute was given. But from the result it can be seen that the recall for the categorie YES churn is still small in comparison to the NO churn.

To solve this imbalance in class, CatBoost model is used. Which is very efficient in solving problems regarding in class imbalance.

CatBoots Model

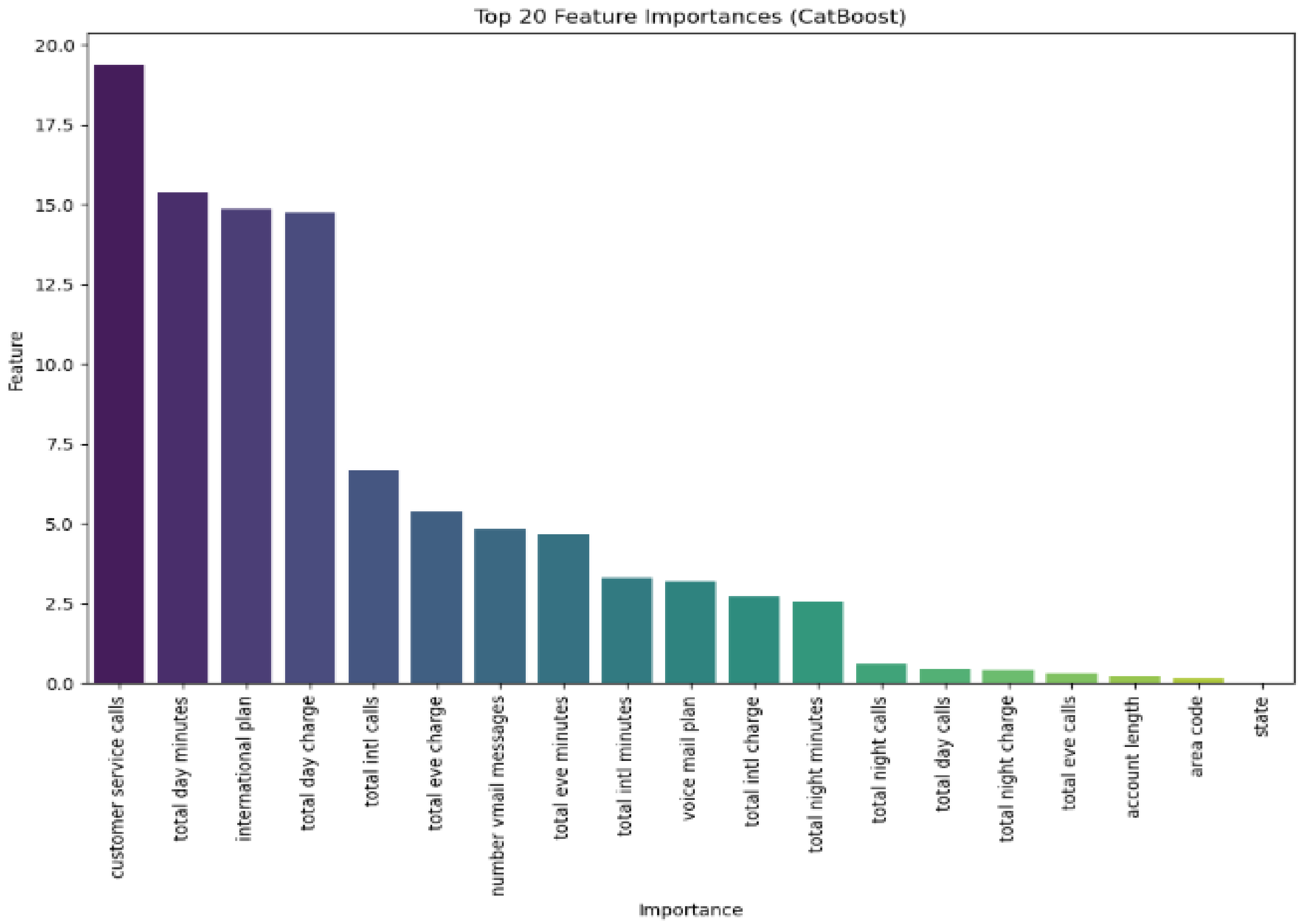




Conclusion:

- it can be seen that the accuracy has improved to 94% and the recall for the 0 is 99% while for the 1 is 64%.
- This means that out of all the models this one performs best, because our interest is in the people who churn, and the recall for them is much improved in comparison to the other models.

Feature importance for CatBoost



Conclusion:the features that are important for the estimation are customer service calls, total day minutes, international plan and total day charge.

Conclusion

- Logistic regression,SVM model,KNN model,Random forest and CatBoost have been tried to predict churn,Random forest and Catboost are considered to be the most suitable models based on the accuracy: Random forest(the accuracy is 92% and the recall of no churn is 99% while for yes churn is 59%). CatBoost(the accuracy is 94% and the recall for the 0 is 99% while for the 1 is 64%). They also handled class imbalance.

- Feature importance was done for this 2 models(Random forest and CatBoost)and the features with the highest importance for Random Forest was customer service calls,total day charge and total day minutes. For the CatBoost was customer service calls,total day minutes,international plan and total day charge.

- While interpreting results of random forest and catboost, as it has been mentioned that some features are more important for the churn modelling. Which means that such features should be considered while dealing with at risk customer, by offering discounts or personalizing plan. Focusing on such features can be useful on devising the plan to reduce churn.

THE END

Any Questions?