

SYRIATEL CUSTOMER CHURN PREDICTION

*Harnessing Data Insights to optimise
Customer Retention*

*By
Evaclaire
Munyika
Wamitu – Lead
Data Scientist*



OVERVIEW

BUSINESS UNDERSTANDING

DATA UNDERSTANDING

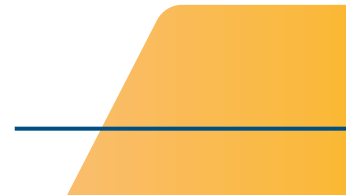
MODELLING

EVALUATION

RECOMMENDATIONS

NEXT STEPS

TABLE OF CONTENTS



PROJECT OVERVIEW

- **Business Challenge:** SyriaTel faces the critical task of identifying customers likely to cancel services, impacting revenue, profitability and brand reputation due to high churn rates.
- **Solution:** Develop a robust predictive model using advanced analytics and machine learning to identify at-risk customers and implement targeted retention strategies, mitigating financial losses from churn.
- **Stakeholder Impact:** Empower SyriaTel executives, marketing teams and customer service reps with data-driven insights to enhance customer satisfaction, loyalty and long-term revenue growth through optimized retention efforts.




BUSINESS UNDERSTANDING

- SyriaTel is a leading telecommunications provider known for its extensive network coverage and innovative services.
 - The company offers a wide range of telecommunication solutions including but not limited to mobile and internet services to both individual and corporate clients.
 - With a commitment to customer satisfaction and technological advancement, SyriaTel strives to maintain its competitive edge in the dynamic telecom industry while facing challenges such as customer retention and market competition.
- 
- 

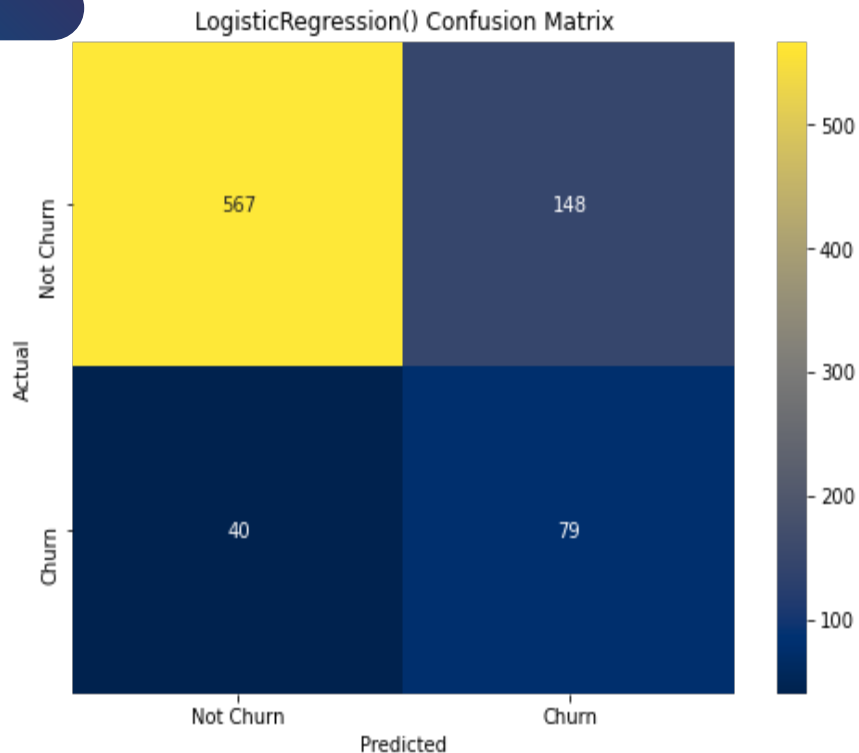


DATA UNDERSTANDING

- The dataset used in this project is called Churn in Telecoms from Kaggle (<https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset>).
 - The dataset provides customer activity data and information about whether they canceled their subscription with the telecom company. The goal is to develop predictive models that can help reduce financial losses caused by customers who do not remain with the company for an extended period.
- 

MODELLING

Logistic Regression Model



- Model correctly predicted 567 instances of class 0 (no churn) and 79 instances of class 1. (churn)
- It misclassified 148 instances of no churn as churn and 40 instances of churn as no churn.

Logistic Regression Model

Train accuracy: 78%

Test accuracy: 77%

Class 0 (majority class)

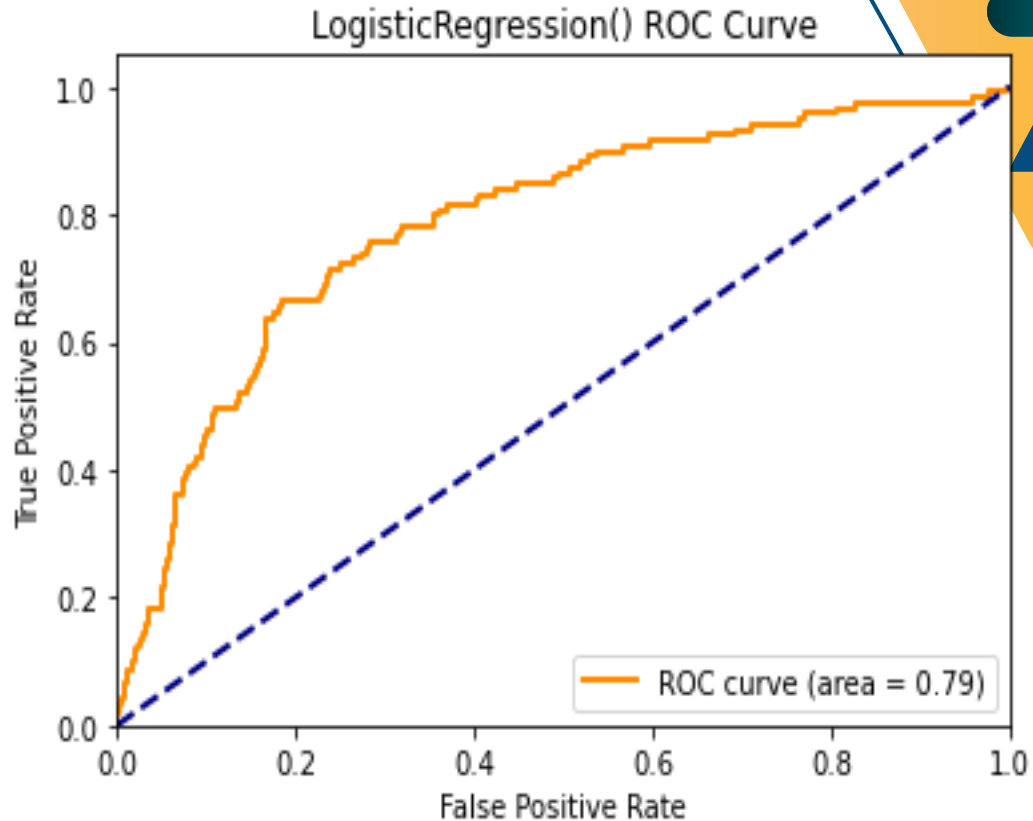
Precision: 93%

Recall/Sensitivity: 79%

Class 1 (minority class)

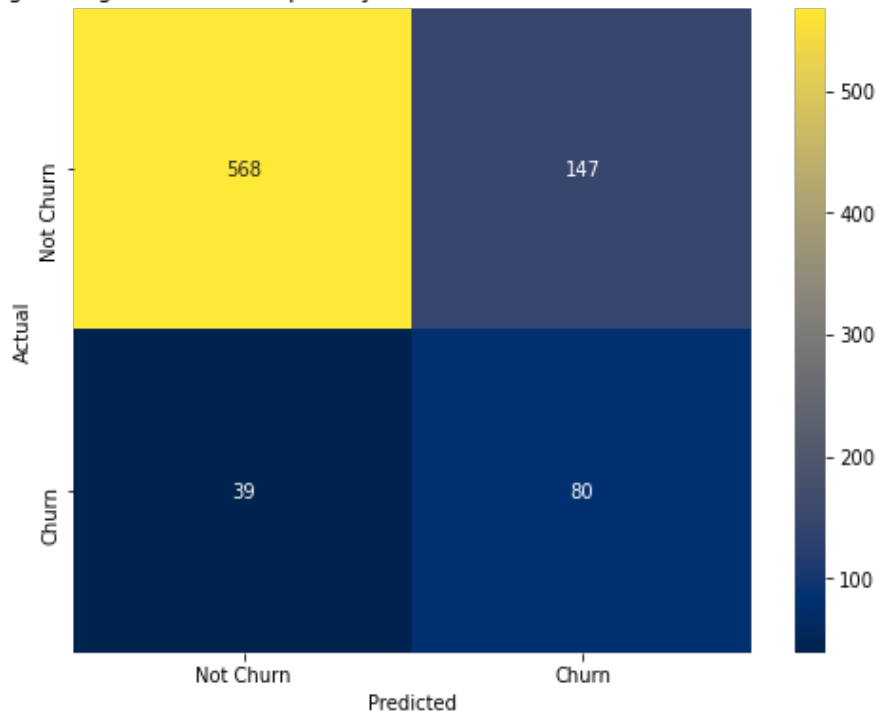
Precision: 35%

Recall: 66%



Logistic Regression (hyperparameter tuned)

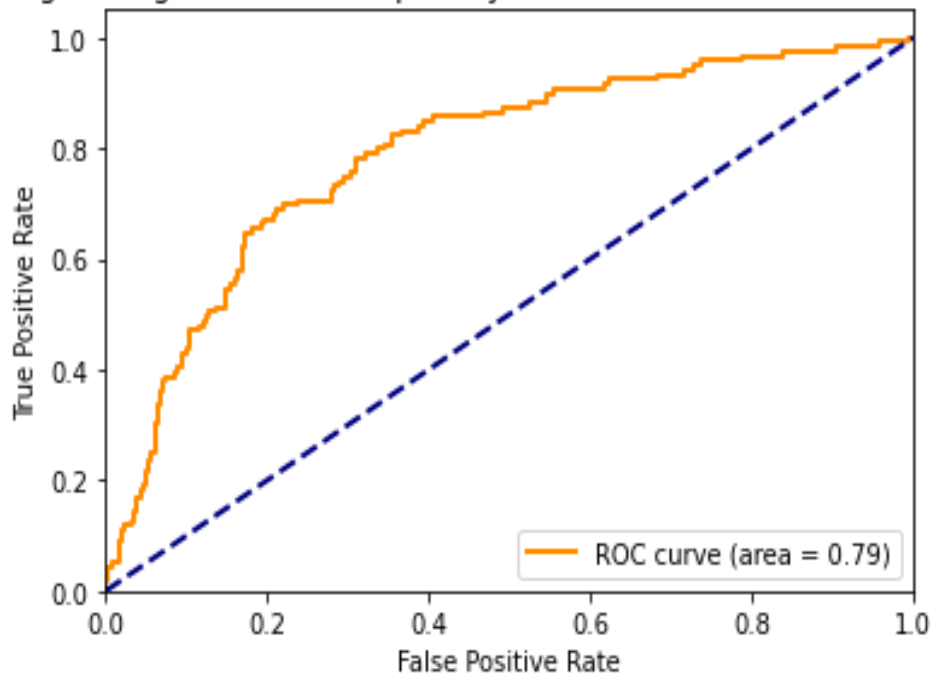
LogisticRegression(C=0.1, penalty='l1', solver='liblinear') Confusion Matrix



- The confusion matrix shows that the model correctly predicted 568 instances of class 0 (no churn) and 80 instances of class 1 (churn)
- It misclassified 147 instances of class 0 as class 1 and 39 instances of class 1 as class 0.

Logistic Regression (hyperparameter tuned)

LogisticRegression(C=0.1, penalty='l1', solver='liblinear') ROC Curve



Train accuracy of 78%
Test accuracy of 78%.

Class 0 (no churn)

Precision: 94%
Recall: 79%

Class 1 (churn)

Precision: 35%
Recall: 67%

Weighted average

F1-score: 0.80

K-Nearest Neighbors (KNN)

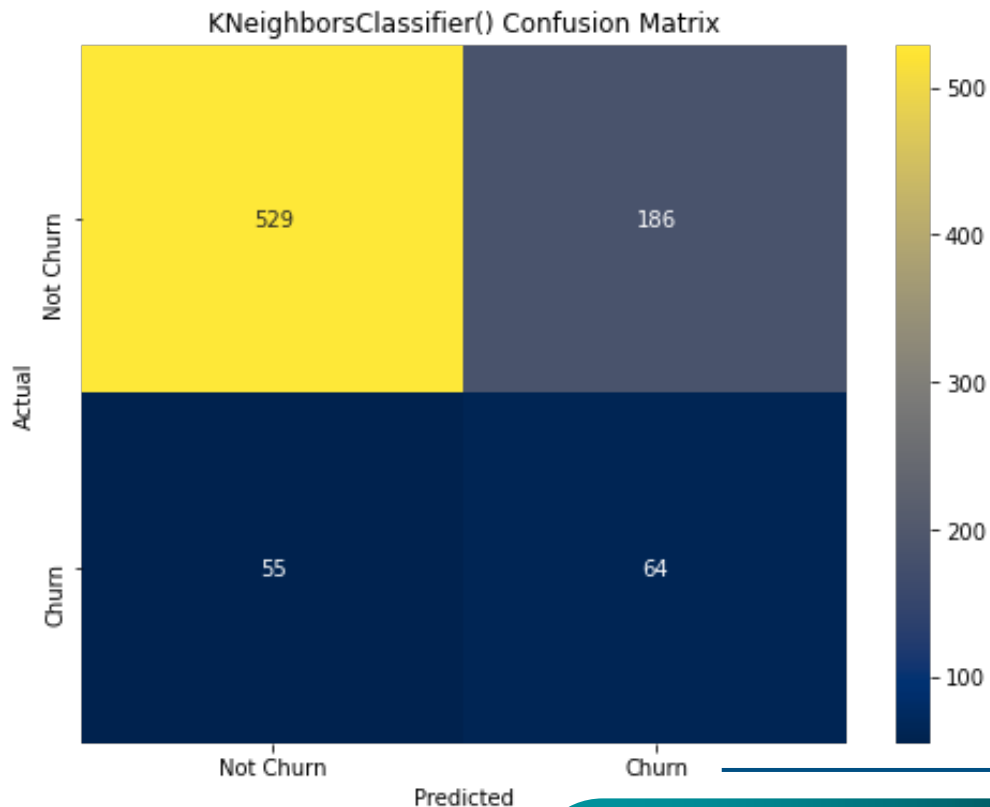
- Confusion Matrix:

Correctly predicted class 0: 529 instances

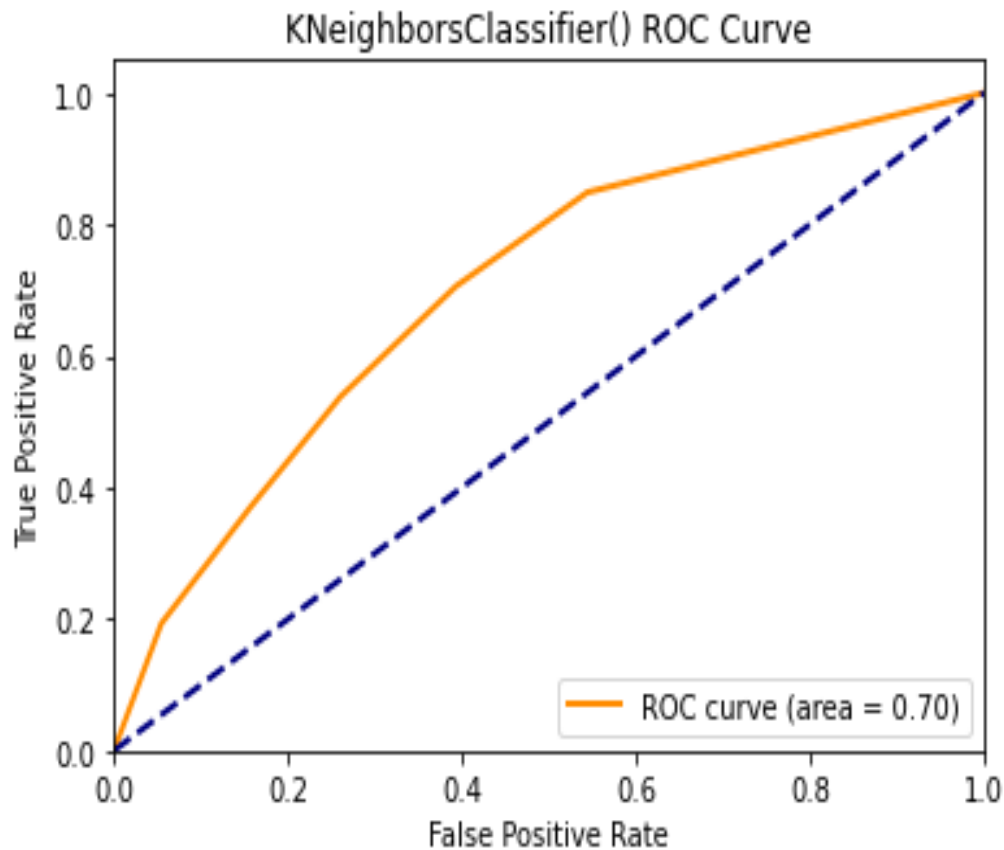
Correctly predicted class 1: 64 instances

Misclassified class 0 as class 1: 186 instances.

Model Comparison: Better accuracy than the hyperparameter tuned logistic regression model but worse precision and recall.



K-Nearest Neighbors (KNN)



Train Accuracy: 84%

Test Accuracy: 71%

Precision (Class 0 - No Churn): 91%

Recall (Class 0 - No Churn): 74%

Precision (Class 1 - Churn): 26%

Recall (Class 1 - Churn): 54%

Weighted Average F1-Score: 0.75

K-Nearest Neighbors (KNN) (hyperparameter tuned)

Best Parameters: $k = 3$, $p = 1$ (Manhattan distance)

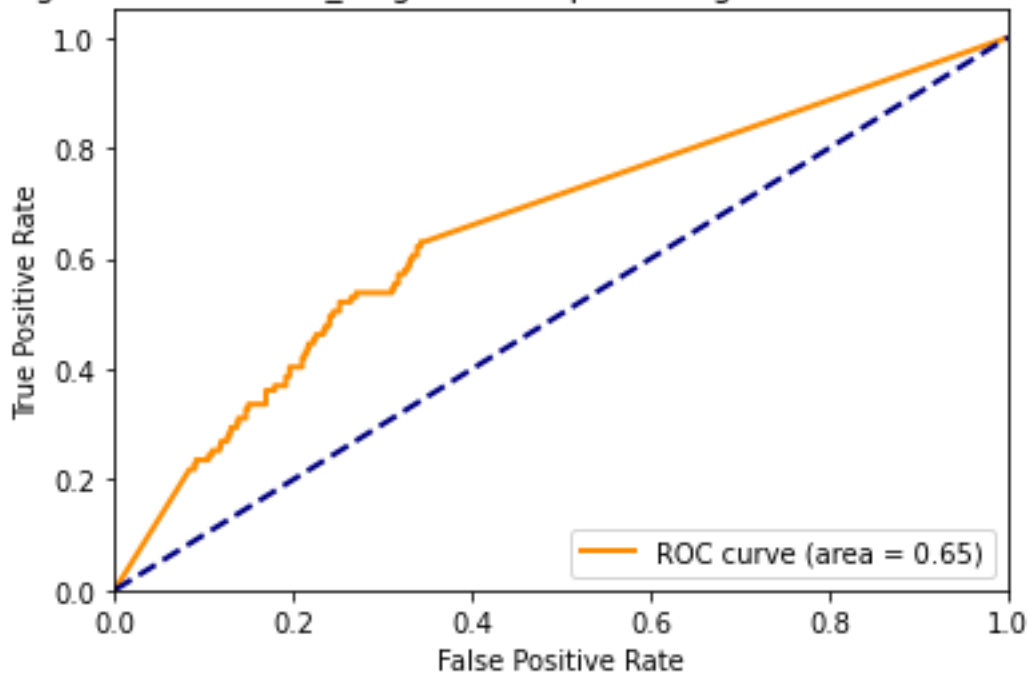
Train Accuracy: 100%

Test Accuracy: 74%

Precision (Class 1 - Churn): Low

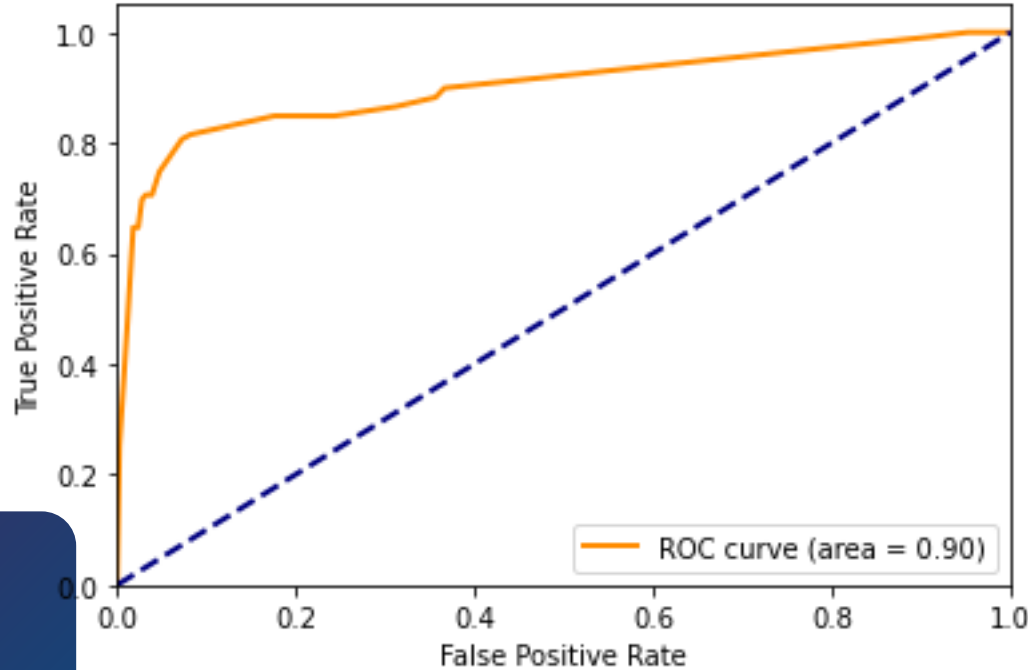
Recall (Class 1 - Churn): Moderate

KNeighborsClassifier($n_neighbors=3$, $p=1$, $weights='distance'$) ROC Curve



Decision Trees

DecisionTreeClassifier(max_depth=6, random_state=1) ROC Curve



Training accuracy: 93%

Test accuracy: 91%

Majority class precision: 97%

Majority class recall: 93%

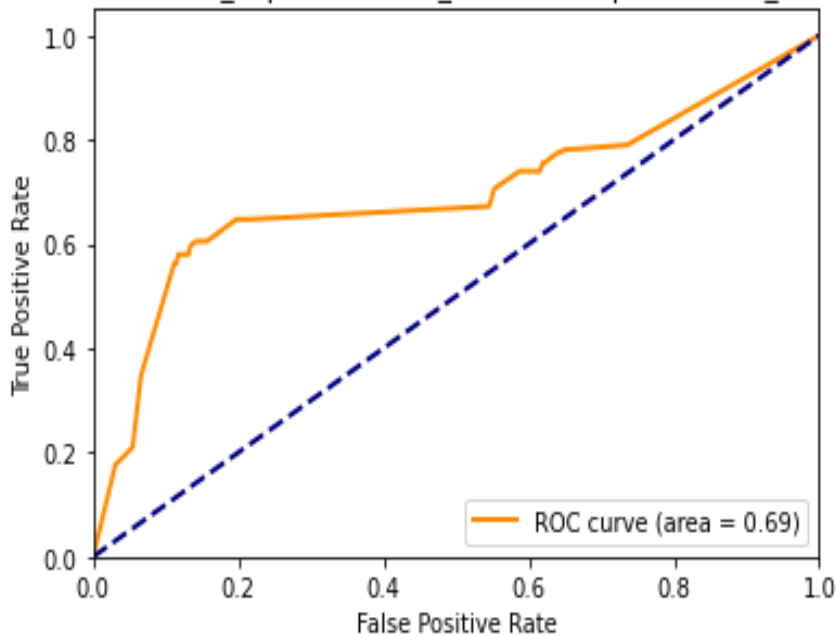
Minority class precision: 65%

Minority class recall: 81%

- The decision tree classifier exhibits robust performance with high accuracies on both training and test sets, effectively distinguishing between classes and achieving balanced metrics.

Decision Trees (Hyperparameter tuned)

DecisionTreeClassifier(max_depth=12, max_features='sqrt', random_state=1) ROC Curve



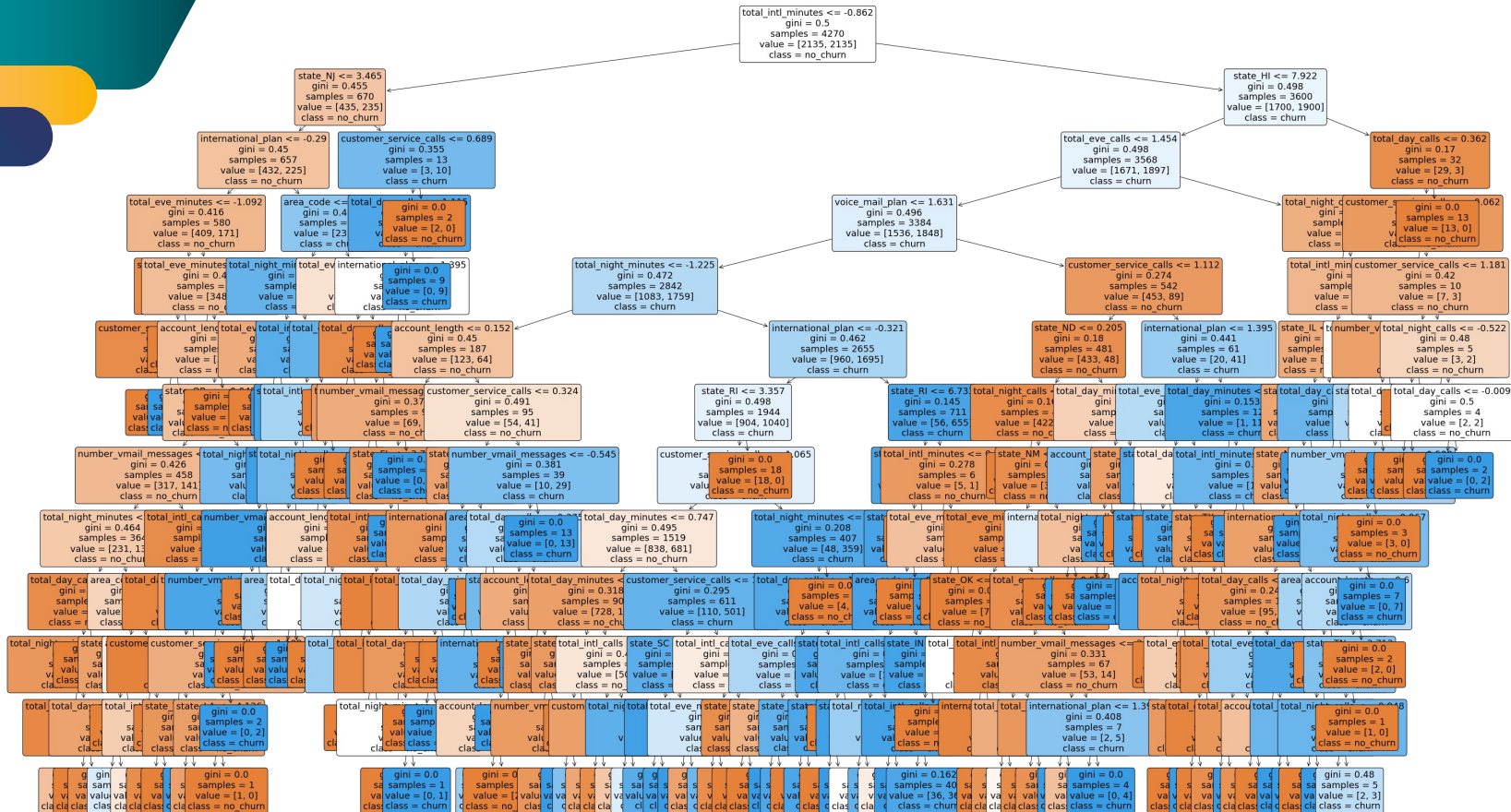
Train accuracy: 89% (down from 93%)

Test accuracy: 82% (down from 91%)

Best Parameters: {'criterion': 'gini',
'max_depth': 12,
'max_features': 'sqrt',
'min_samples_leaf': 1,
'min_samples_split': 2}

- Hyperparameter tuning may at times lead to overfitting.

Decision Trees (Illustrated)



Random Forest Classifier

Training accuracy: 100%

Test accuracy: 93%

Precision: 0.95 (class 0), 0.75 (class 1)

Recall: 0.96 (class 0), 0.71 (class 1)

F1-score: 0.96 (class 0), 0.73 (class 1)

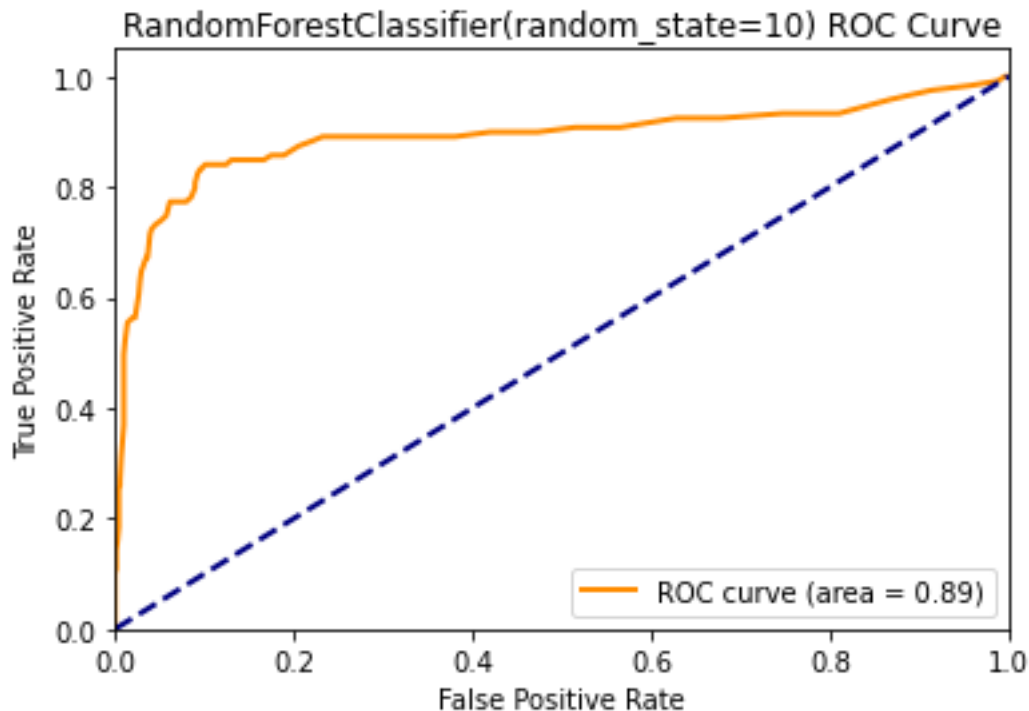
Confusion matrix:

687 correct (class 0)

85 correct (class 1)

28 misclassified (class 0)

34 misclassified (class 1)



- The model shows excellent generalization with high accuracy and robust performance across classes.

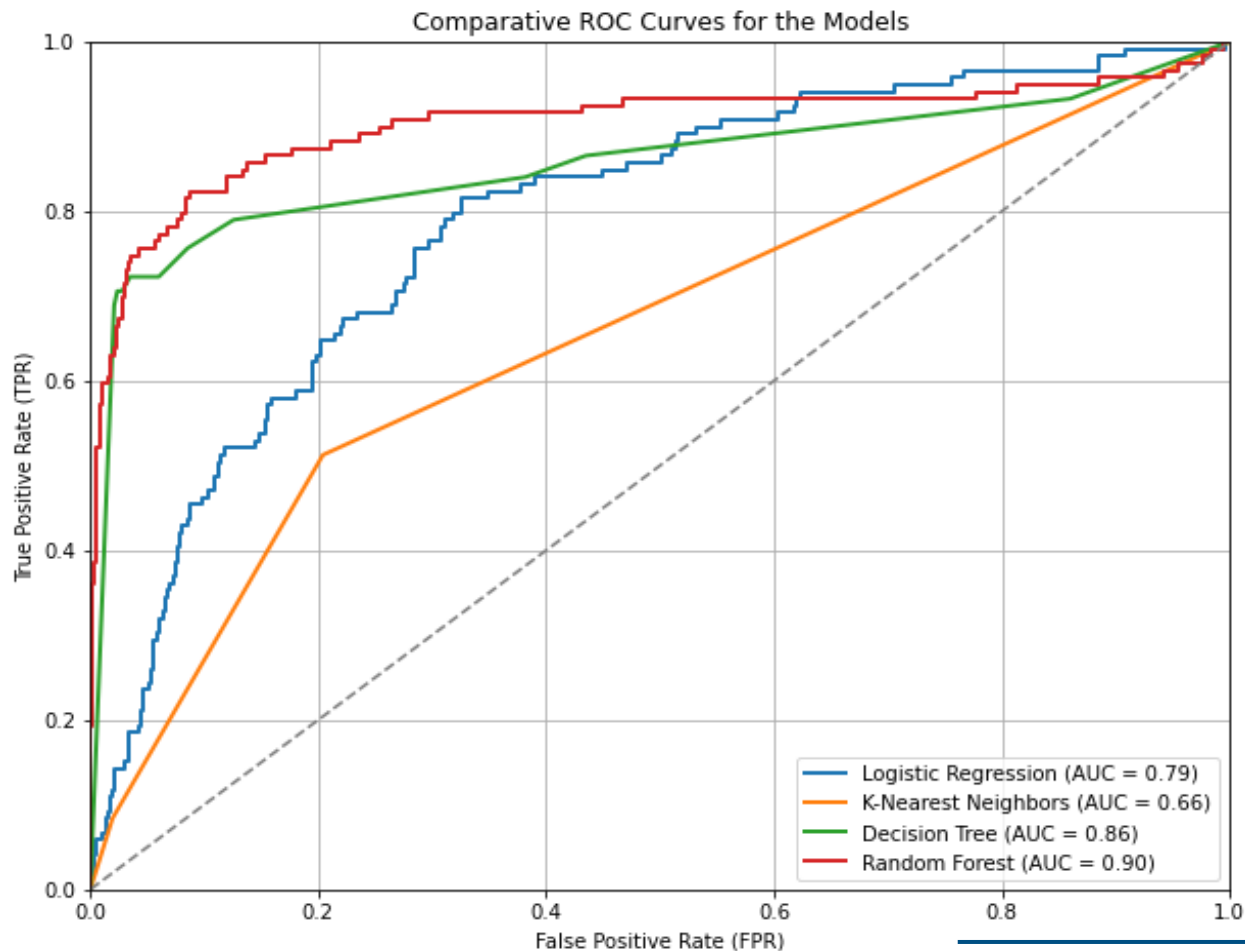


EVALUATION

Random Forest model has the highest AUC (0.90), indicating the best performance.

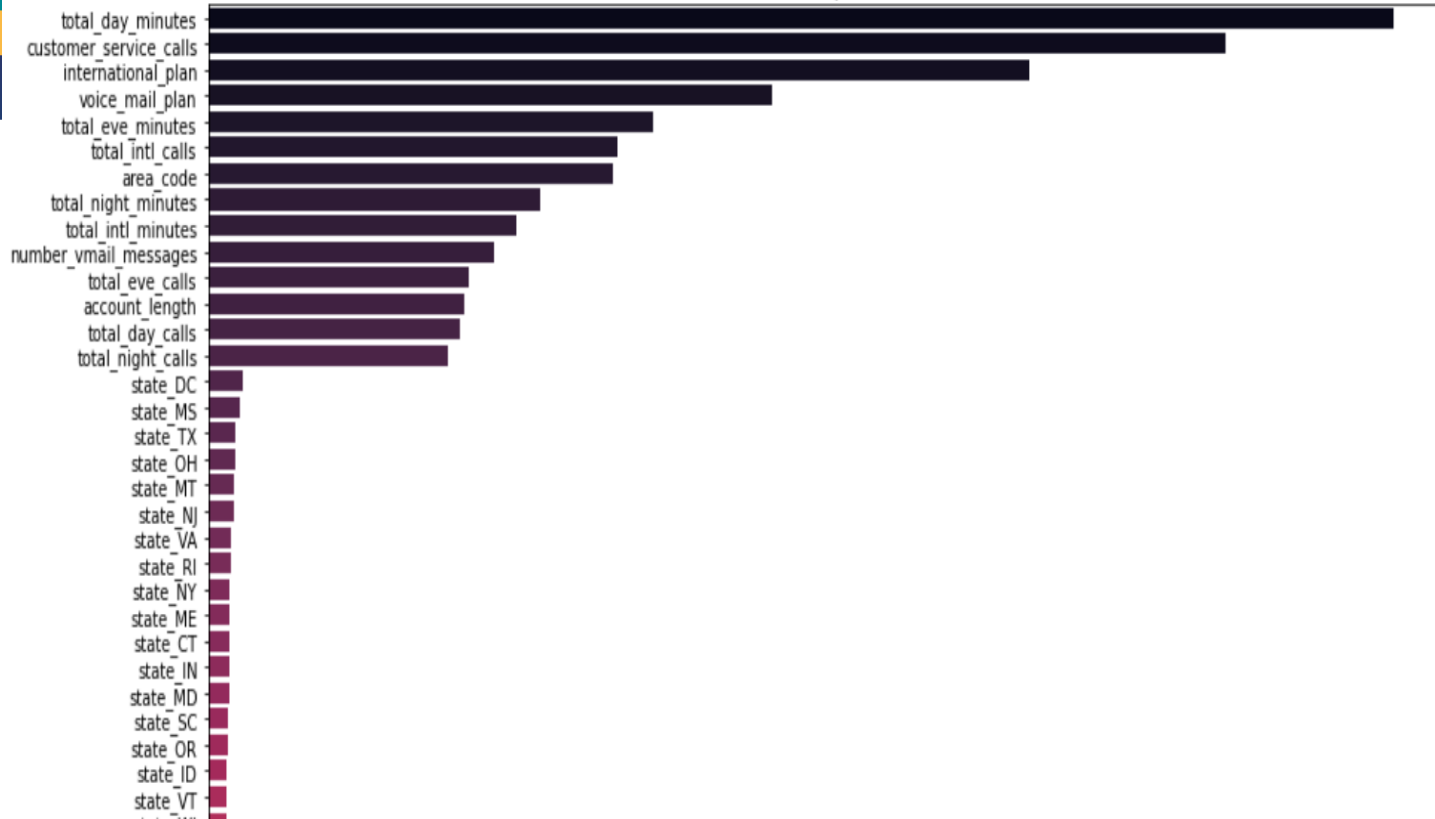
Decision Tree follows with an AUC of 0.86, showing strong predictive capability.

Logistic Regression and K-Nearest Neighbors have lower AUCs of 0.79 and 0.66, respectively, indicating less effectiveness.



Feature Selection

Feature Importances



Conclusions

Random Forest Superiority:

Random Forest Classifier emerges as the top performer, achieving 100% training accuracy and 93% test accuracy, showcasing exceptional generalization without overfitting or underfitting.

Call Usage Dominance:

Call usage features like total day minutes and number of customer service calls dominate as top predictors of churn, emphasizing the significance of customer calling behaviour in churn prediction.

Feature Importance Insights:

Customer tenure, service characteristics, and geographical factors play substantial roles in predicting churn, while state-level features demonstrate lower influence. Consideration of feature selection and model optimization is crucial for further enhancing predictive accuracy and interpretability.

RECOMMENDATIONS

Recommendations

1. **Recommendation:** Utilize Random Forest Classifier for churn analysis due to exceptional performance across metrics.
2. **Call Usage Monitoring:** Monitor call patterns for proactive retention strategies.
3. **Customer Satisfaction:** Enhance satisfaction with loyalty programs to deter churn.
4. **Regional Analysis:** Tailor strategies based on regional churn differences.
5. **Feature Prioritization:** Prioritize key features over state-level features.
6. **Model Enhancement:** Consider techniques like feature elimination and explore XGBoost or SVM-support vector machines.
7. **Conclusion:** Focus on critical features to improve retention and revenue.

NEXT STEPS

Next Steps

Cross-Functional Collaboration:

- Foster collaboration between data scientists, marketing and customer service teams.
- Ensure alignment of strategies and efforts towards reducing churn and maximizing customer lifetime value.

Regular Performance Evaluation:

- Establish a schedule for regular evaluation of the model's performance.
- Use feedback to fine-tune strategies and improve accuracy over time.

Call Usage Monitoring System:

- Develop a system to monitor call usage patterns in real-time.
- Implement alerts for significant changes in calling behaviour.
- Establish protocols for proactive customer outreach based on identified patterns.



THANK YOU!

For any queries feel free to reach me at
evamunyika@gmail.com