TELECOM CHURN CASE STUDY

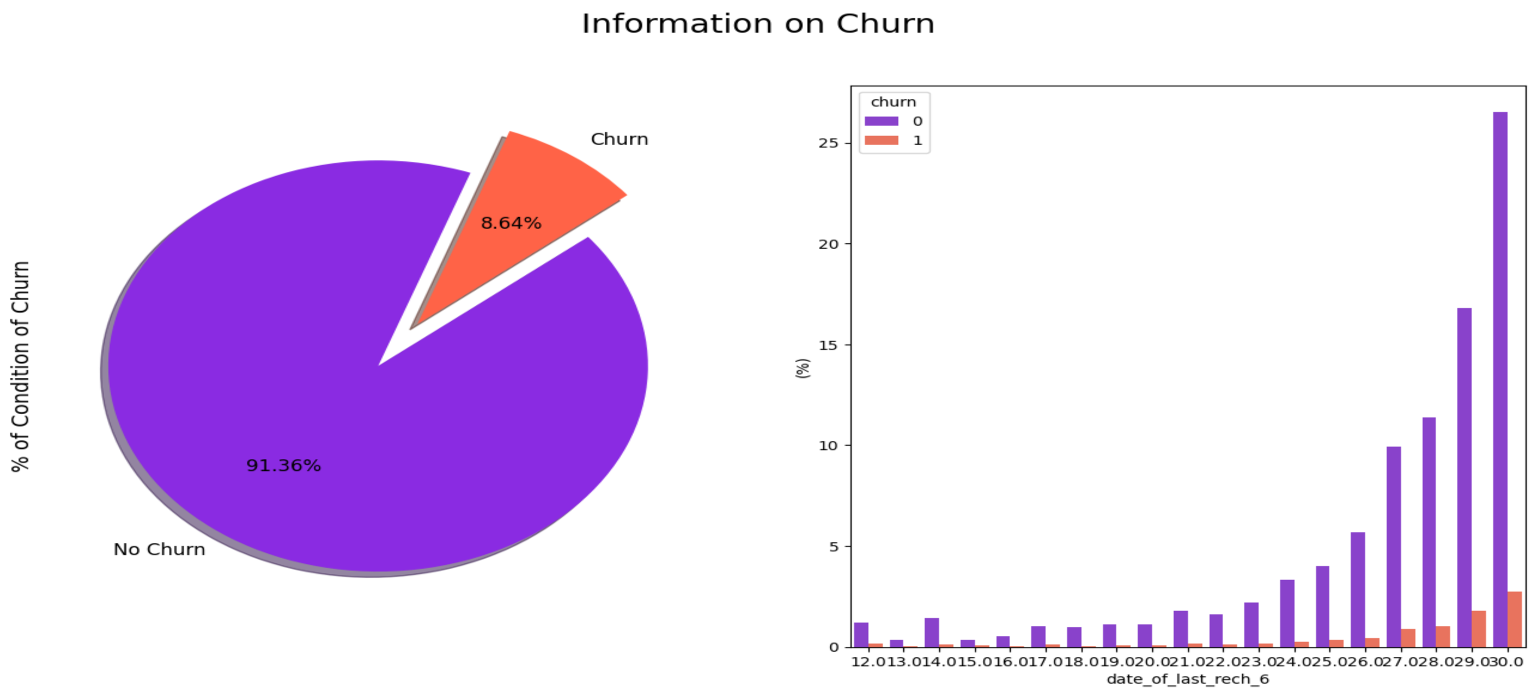
Submitted By:

Harshika Singh

PROBLEM STATEMENT

In this project, you will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

INFORMATION ON CHURN



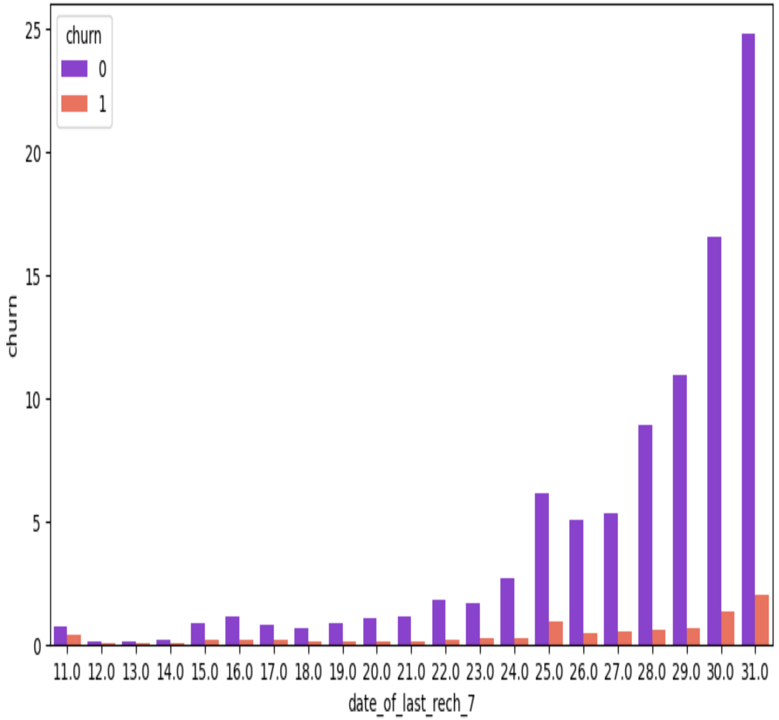
Insights:

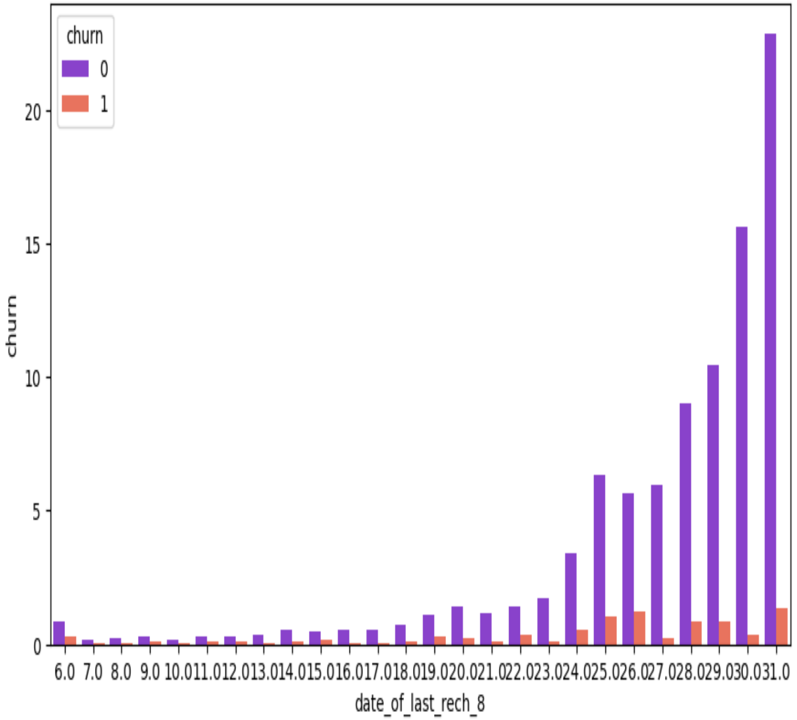
1. 91.36% of customers do not churn, while 8.64% do.

2. Highlights the imbalance in churn data, crucial for model evaluation and selection.

3. Month 6: Churn is minimal in the initial days and peaks near the month's end.

Date of Last Recharge and Churn





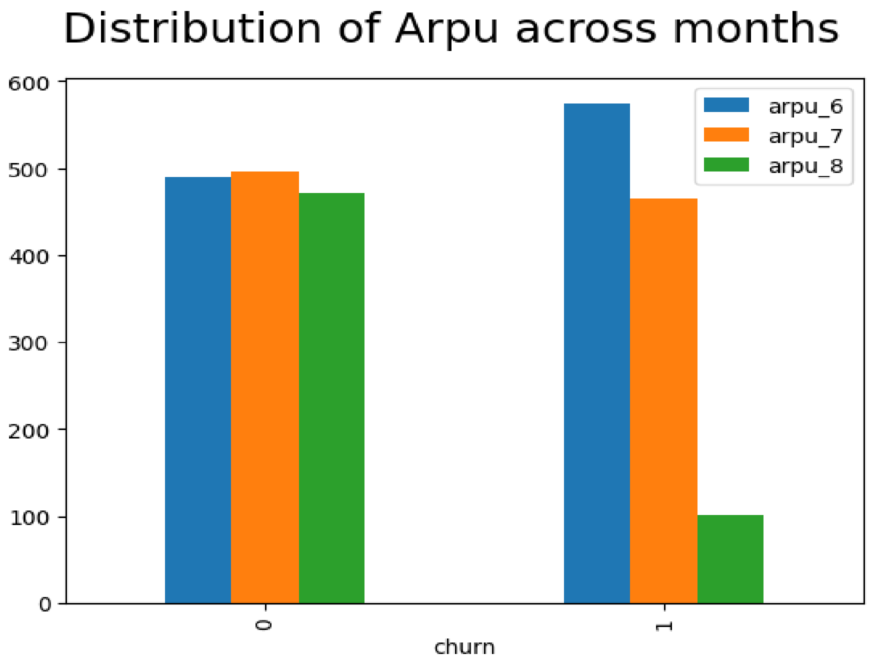
Insights:

Month 7: Similar trend as month 6, with a spike near the 31st.

Month 8: Consistent pattern; churn increases towards the end of the month.

Indicates the need to focus on behavior patterns towards month-end.

Distribution Of Arpu Across Months



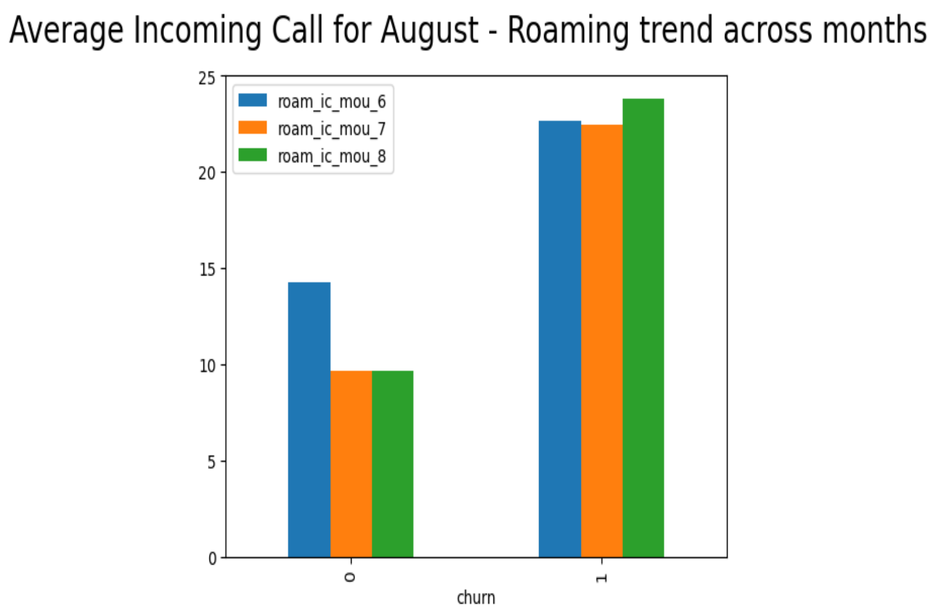
Insights:

1. Non-churned Users (Churn = 0): The ARPU remains relatively stable across months.

2. Churned Users (Churn = 1): There is a noticeable decrease in ARPU, especially in month 8.

3. Implication: Lower ARPU in the last observed month may be a strong indicator of potential churn, suggesting reduced usage before departure.

Average Incoming Call for August - Roaming trend across months



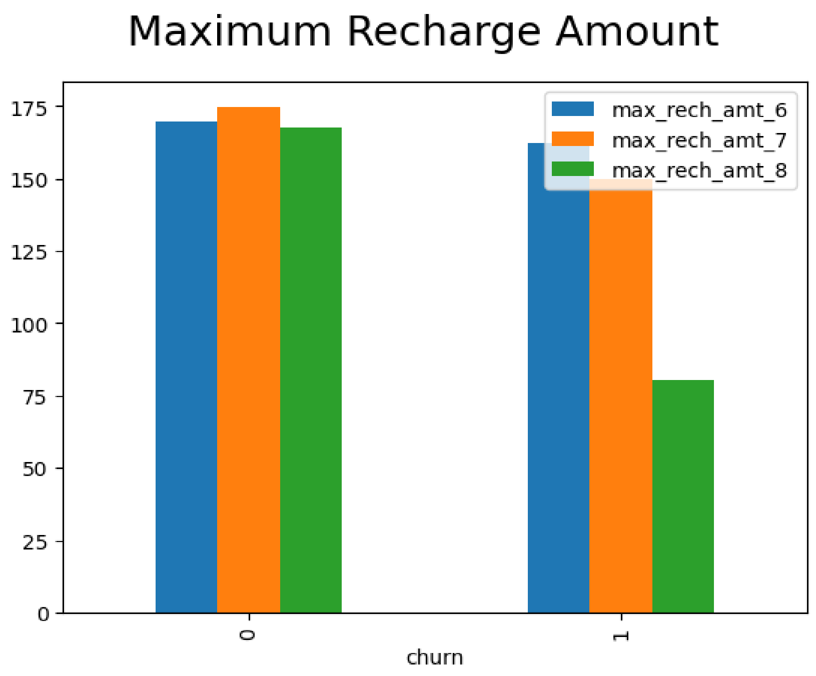
Insights:

1. Churned customers have higher average incoming roaming call minutes in July and August than non-churned customers.

2. Both churned and non-churned customers show a month-over-month increase in roaming call minutes.

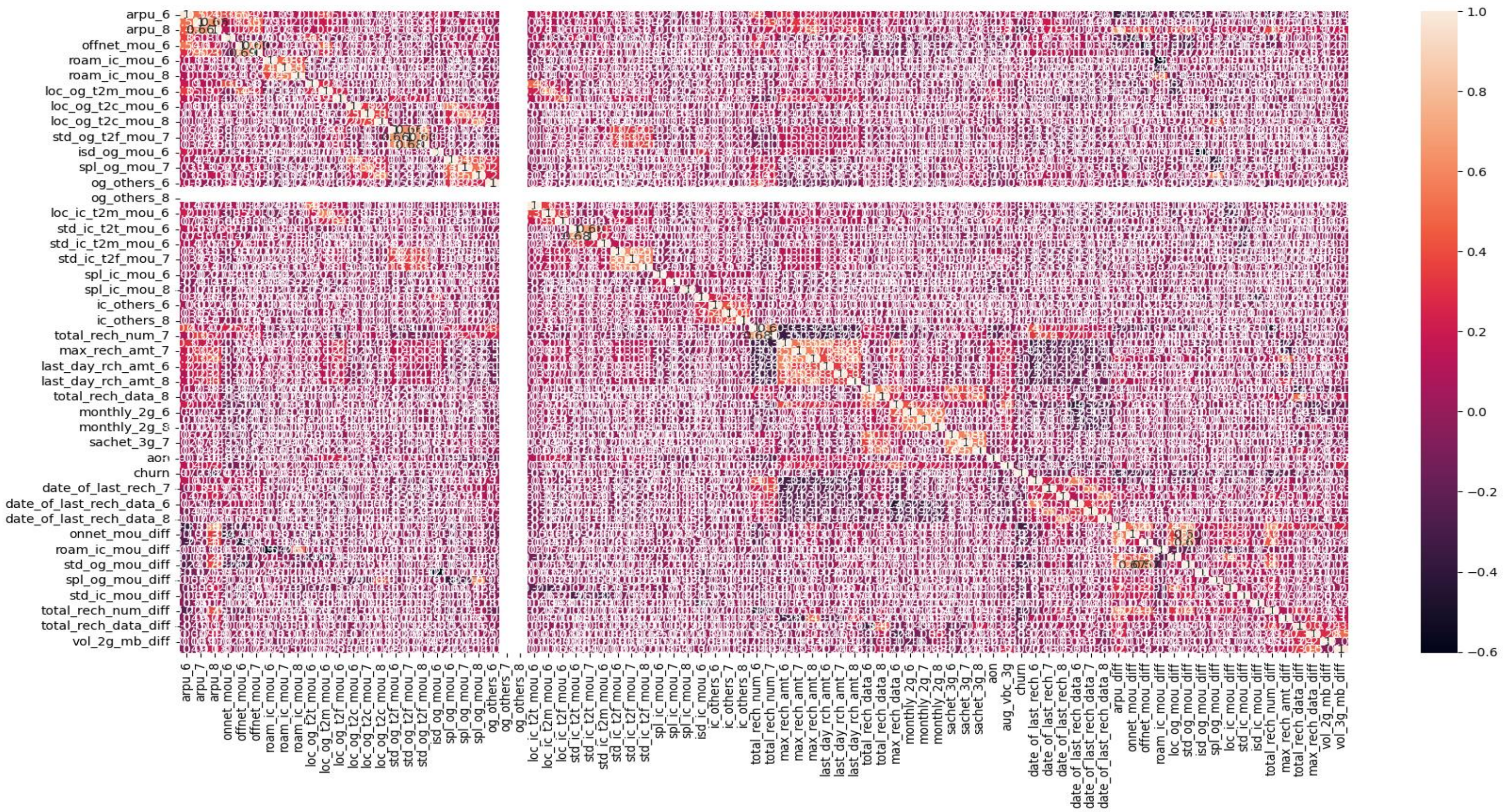
3. Roaming usage may be a churn indicator, suggesting the need for further analysis and targeted retention strategies.

Maximum Recharge Amount Analysis



Insights:
1. Non-churned Users (Churn = 0): Recharge amounts remain relatively stable across months.
2. Churned Users (Churn = 1): A significant drop in recharge amount is observed in month 8.
3. Implication: This decline could indicate that users planning to churn reduce their recharge amount prior to leaving.

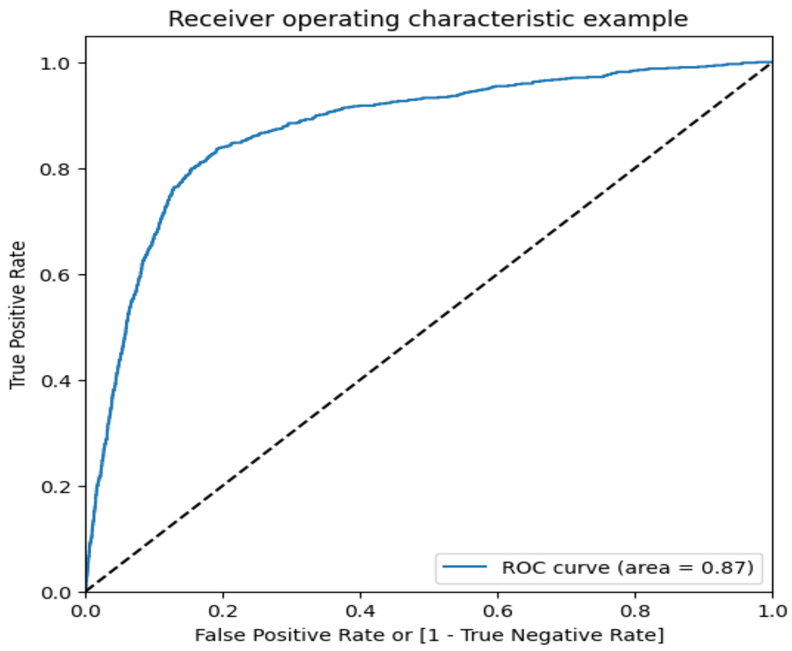
Scaling using MinMax Scaler

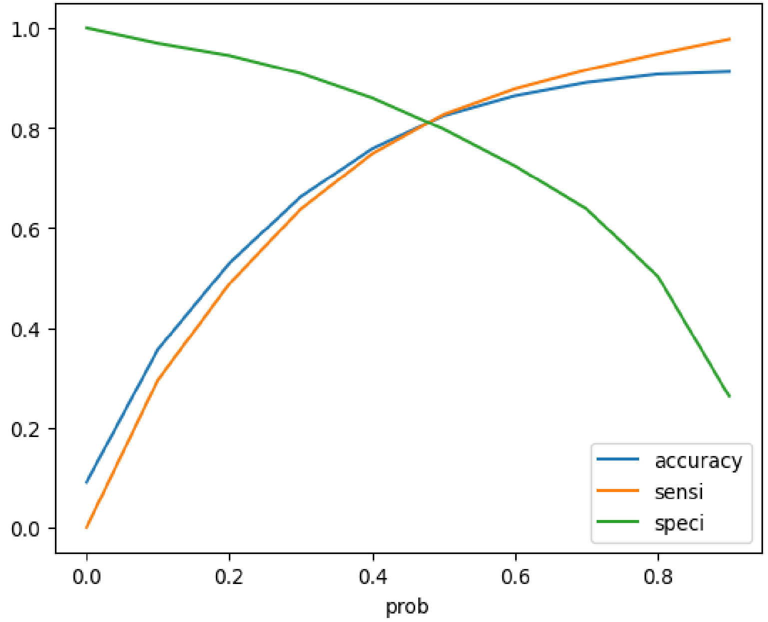


Insights:

The heatmap shows strong correlations between usage patterns and recharge behavior, with churn linked to factors like 'date of last recharge,' emphasizing the role of customer engagement. The mix of positive and negative correlations highlights the complexity of churn drivers and the need for a comprehensive understanding of customer behavior.

Creating Logistic Regression model with RFE





Insights:

1. AUC = 0.88 indicates strong model performance.

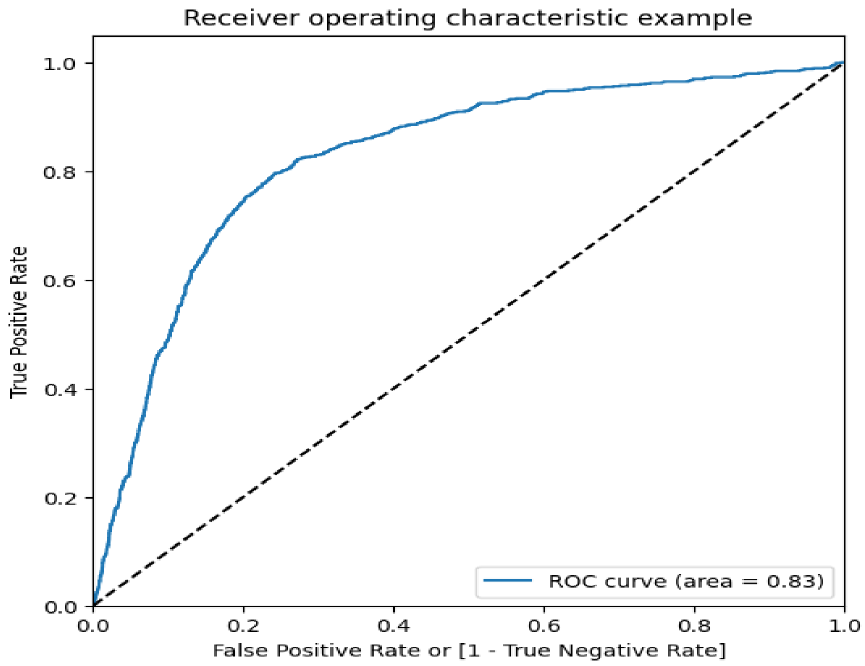
2. High true positive rate, low false positive rate.

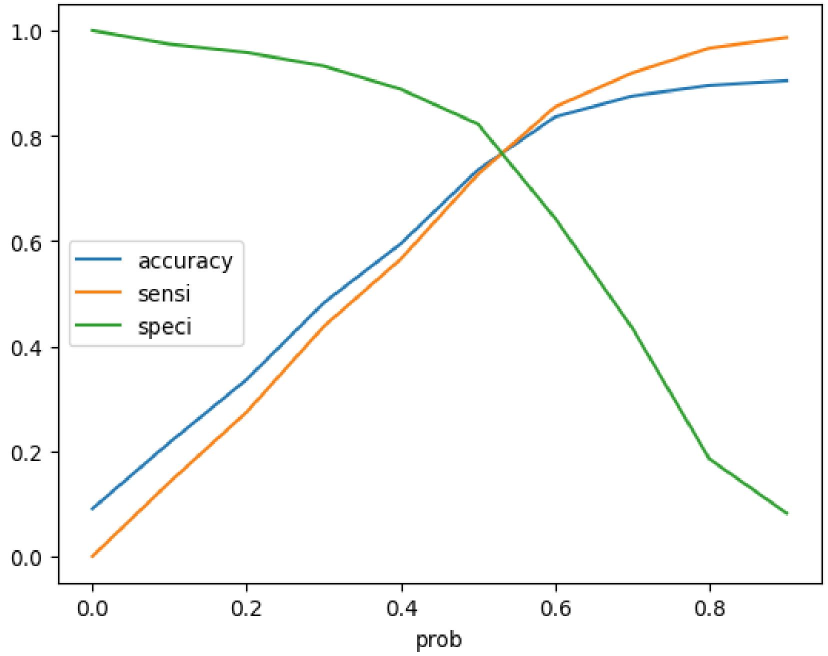
3. Shows accuracy, sensitivity, and specificity against probability thresholds.

4. Sensitivity and specificity have an inverse relationship.

5. Optimal threshold balances sensitivity and specificity.

Recreating the Logistic model again with uncorrelated Features





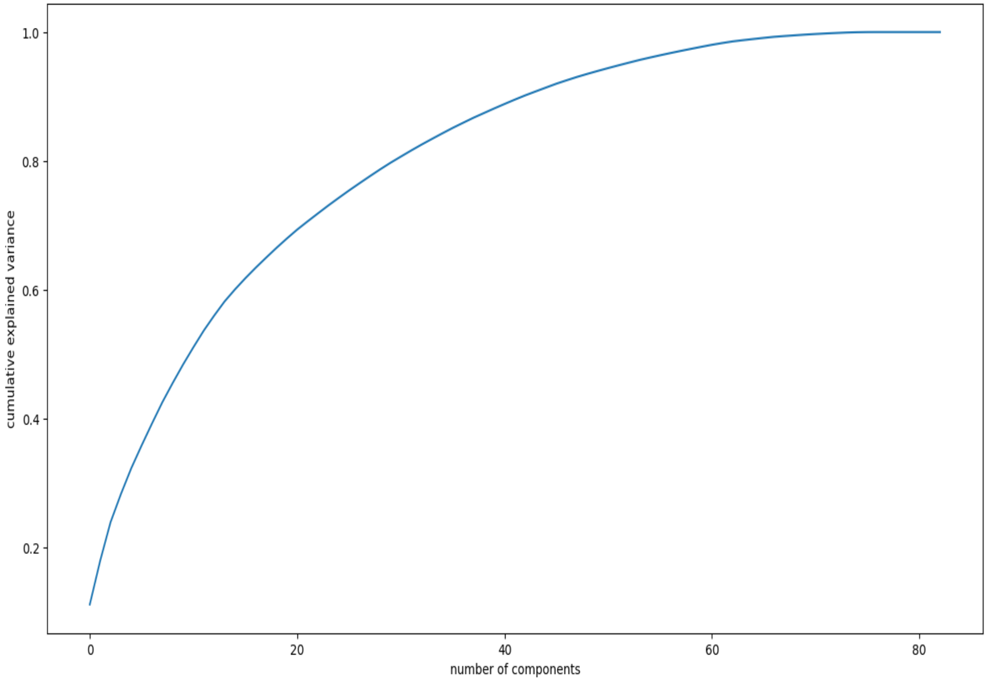
Insights:

1. AUC = 0.83 shows good performance but lower than AUC = 0.88. Model has room for improvement.

2. The graph shows the relationship between probability and accuracy, sensitivity, and specificity.

3. As the probability increases, sensitivity and accuracy increase, while specificity decreases.

Feature Reduction With PCA



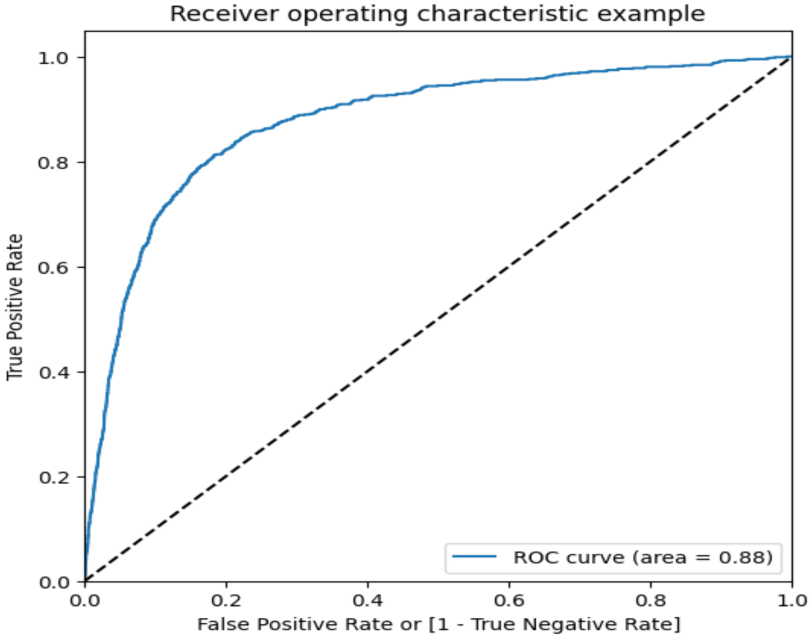
Insights:

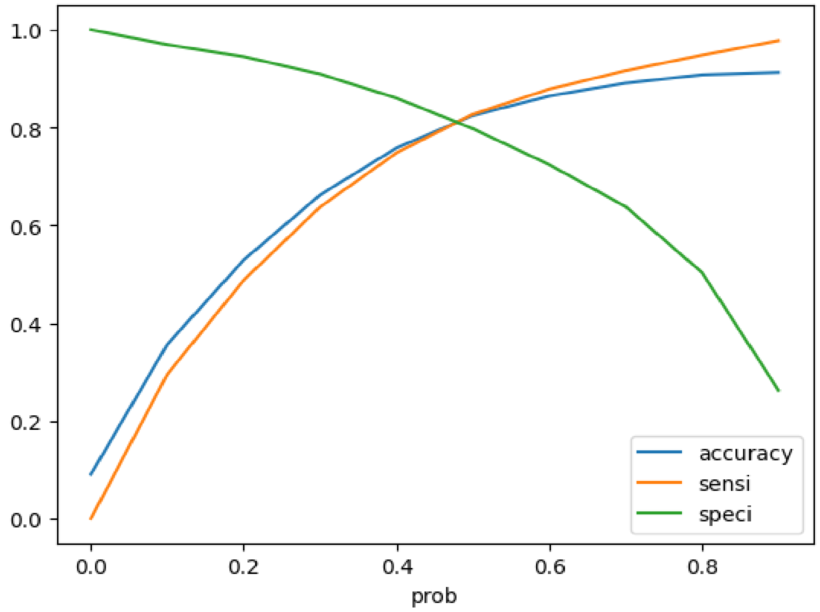
1.The plot shows the cumulative explained variance as a function of the number of principal components.

2.The curve indicates that with increasing components, the explained variance also increases.

3.The curve plateaus around 60 components, suggesting that adding more components beyond that point provides diminishing returns in terms of explained variance.

Logistic Regression With PCA





Insights:

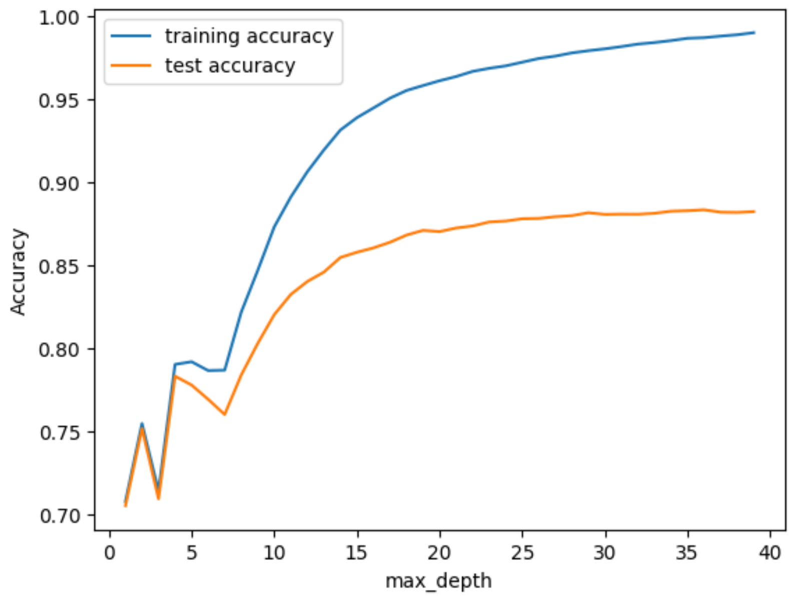
1. The ROC curve shows the trade-off between sensitivity and the false positive rate. An AUC of 0.88 indicates good model performance in distinguishing between classes.

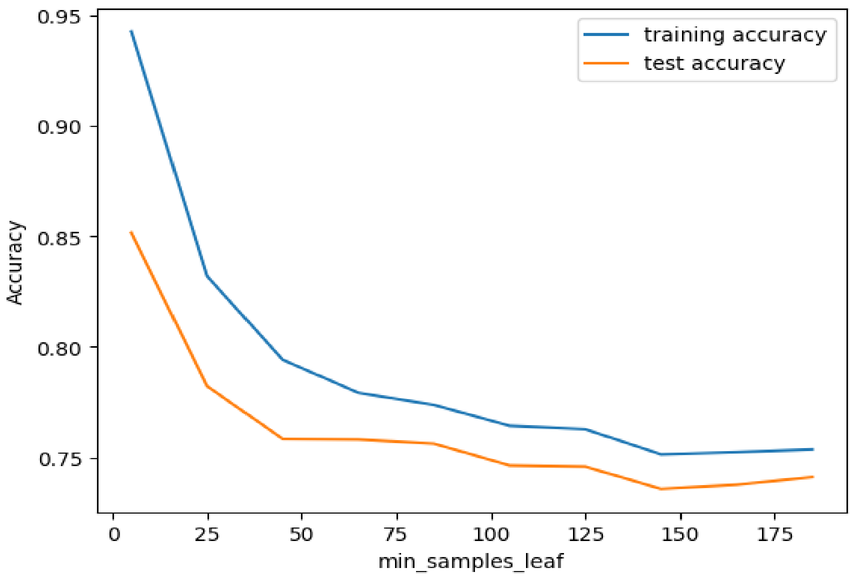
2. The model shows a trade-off between sensitivity and specificity: higher thresholds increase sensitivity but reduce specificity.

3. The accuracy remains high and stable across thresholds, indicating robust performance.

4. The optimal threshold depends on priorities: use a lower threshold to capture positives or a higher one to minimize false positives.

Decision Tree Classifier + PCA with Hyperparameter Tuning



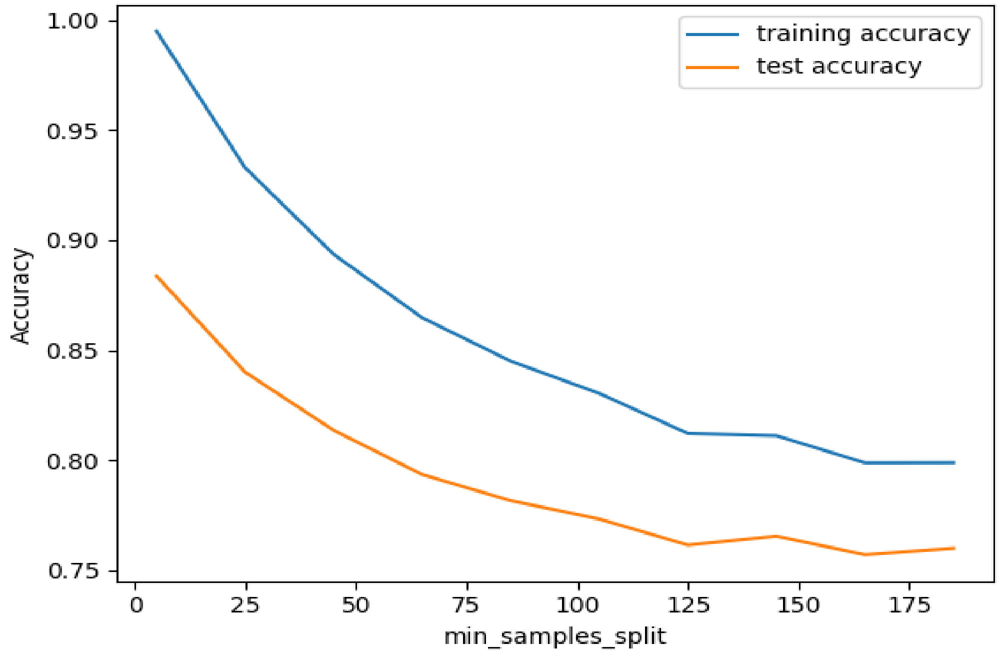


Insights:

1. The graph shows training accuracy rising with tree depth, while test accuracy plateaus around depth 20, suggesting overfitting at greater depths.

2. Higher training accuracy than test accuracy suggests potential overfitting.

3. Increasing min_samples_leaf reduces both accuracies, balancing complexity and generalization. Lower values favor training accuracy, while higher values improve generalization.



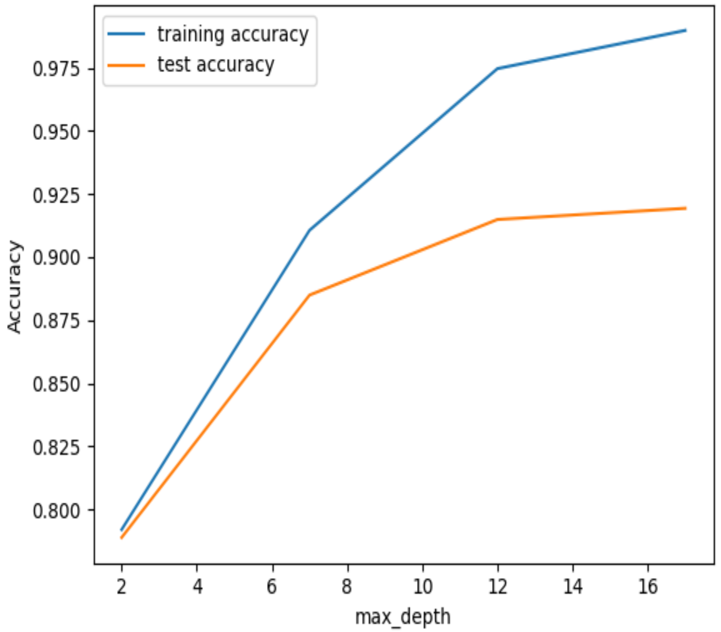
Insights:

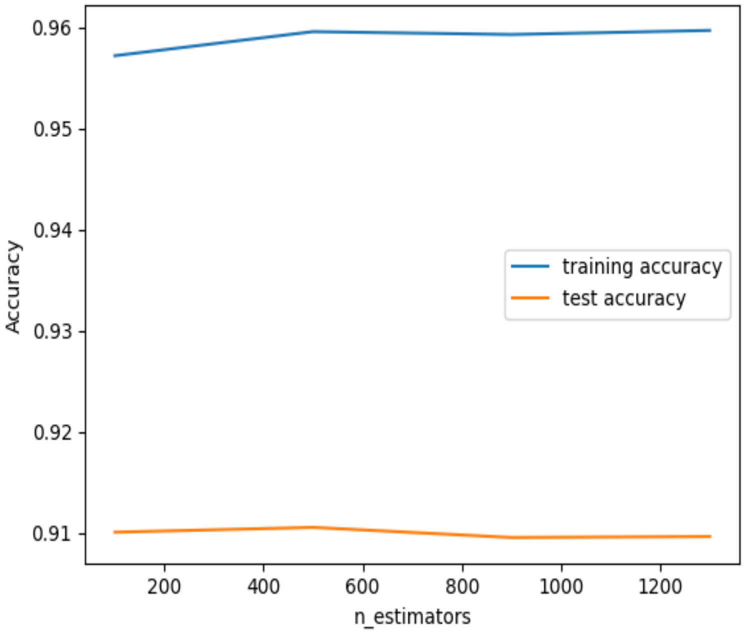
1. Overfitting: The training accuracy is significantly higher than the test accuracy, indicating overfitting.

2. Min_samples_split: As the minimum number of samples required to split an internal node increases, both the training and test accuracy decrease.

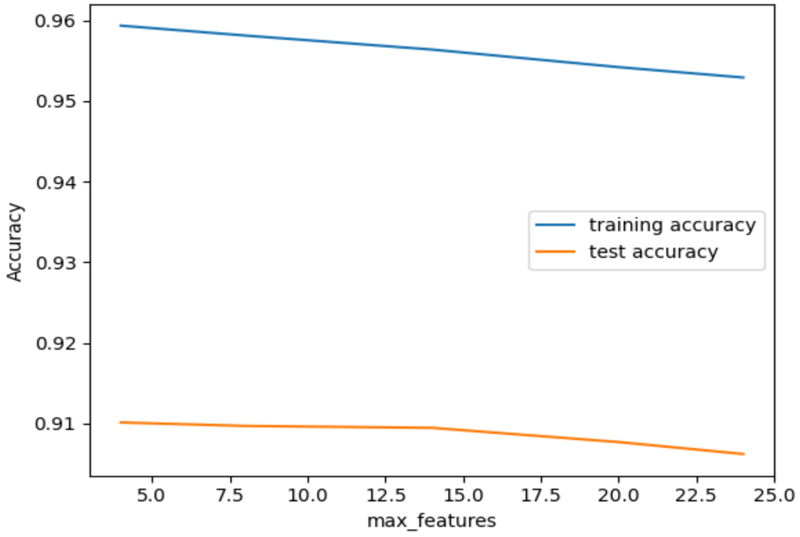
3. Optimal Value: The optimal value for min_samples_split appears to be around 125, where the test accuracy is relatively high and overfitting is minimized.

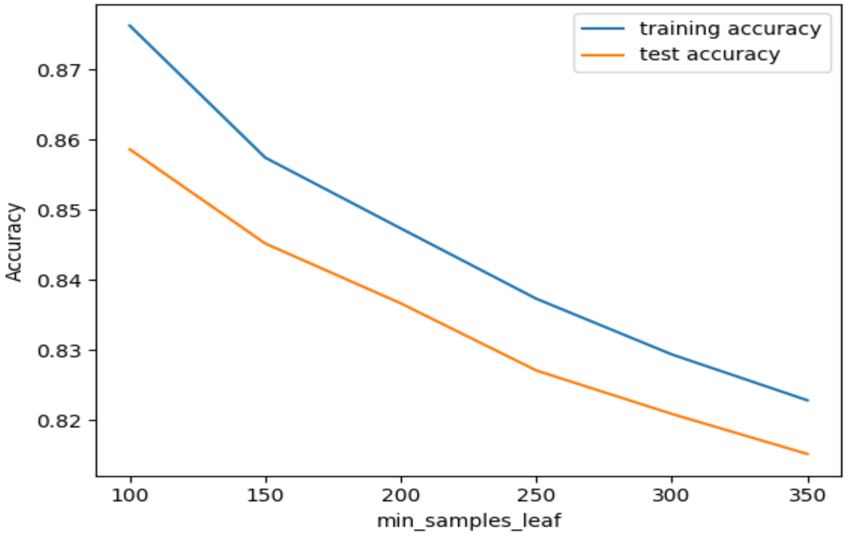
Random Forest Classification + PCA with Hyperparameter tuning





Insights:
1. Training accuracy rises with depth, reflecting more complex pattern learning.
2. Test accuracy plateaus around depth 12, indicating potential overfitting beyond this point.
3. Training accuracy slightly increases with more estimators, while test accuracy remains constant, indicating overfitting.





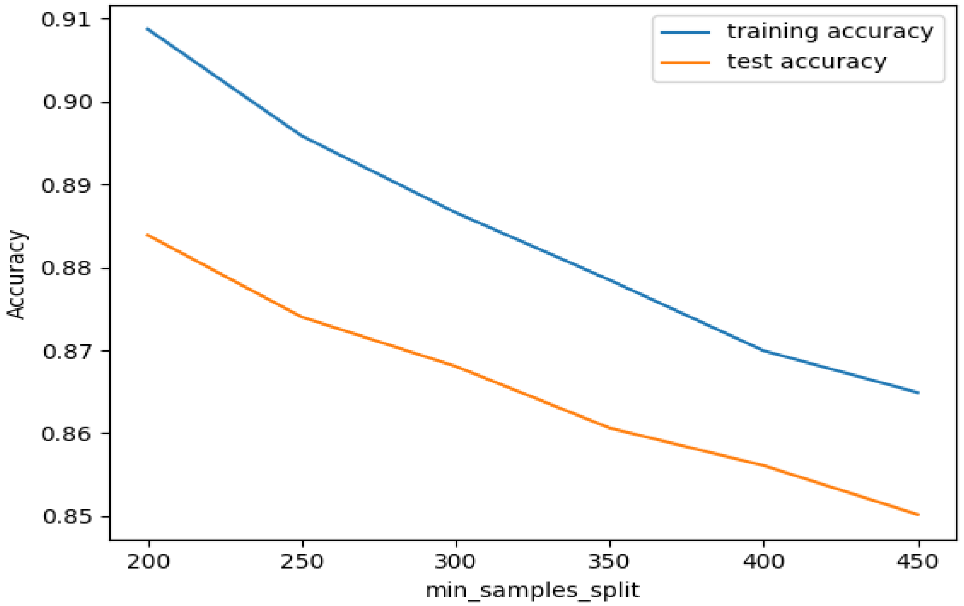
Insights:

1. The training accuracy is higher than test accuracy, this indicates potential overfitting.

2. The test accuracy decreases as max_features increases. This indicates that the model performs better with fewer features.

3. Higher training accuracy than test accuracy suggests overfitting.

Increasing min_samples_leaf lowers both accuracies, favoring smaller values for this model.



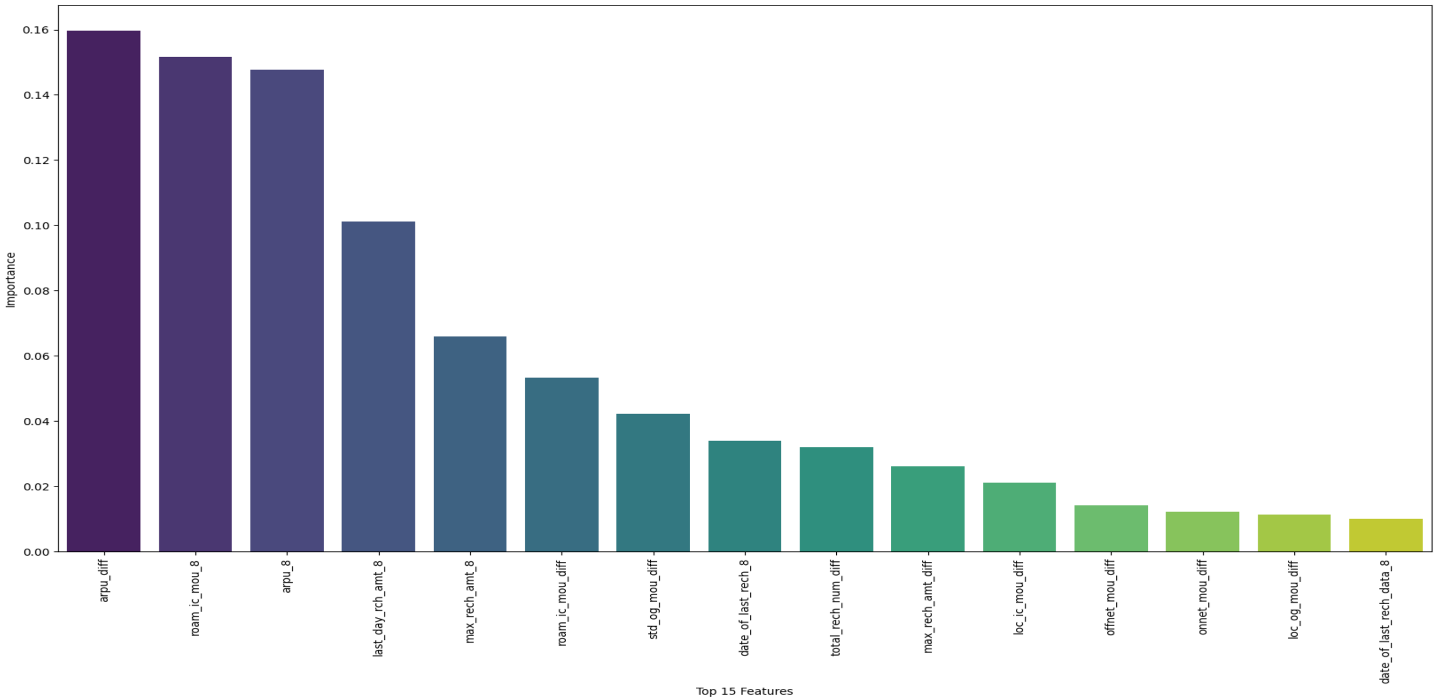
Insights:

1. Training accuracy exceeds test accuracy for all min_samples_split values, indicating overfitting and poor generalization.

2. Increasing min_samples_split reduces both training and test accuracy, simplifying the model with fewer tree nodes.

3. The widening gap between training and test accuracy as min_samples_split increases further highlights overfitting.

Feature Importance of the Final Random Forest Model



Insights:

1. **Arpu_diff is the most important feature in predicting churn.**
2. **Roaming income and outgoing calls are also crucial factors.**
3. **Recharge amount (both last day and maximum) plays a significant role.**
4. **Date of last recharge and total recharge numbers are also key indicators.**

Important Churn Indicators & Actionable Recommendations

1. ARPU in August: Low ARPU suggests higher churn risk.
2. Roaming Incoming Calls in August: Roaming usage indicates potential churn, as the number may no longer be in use.
3. Last and Maximum Recharge Amount in August: A decrease in recharge amount signals a higher likelihood of churn.
4. STD Outgoing Calls (June-July vs. August): A rising difference in STD calls points to churn risk.
5. Recharge Amount Difference (June-July vs. August): A widening gap in recharge amounts suggests increased churn risk.
6. On-Net Calls Difference (June-July vs. August): A growing gap in on-net calls indicates potential churn.

THANK YOU