Customer Churn Prediction

➢ Based on the analysis and model building we've done:

Data Preprocessing:

- We successfully loaded the dataset and found no missing values.
- Categorical features were identified and converted to numerical format using one-hot encoding.
- Class imbalance in the target variable ('Churn') was addressed using SMOTE, resulting in a balanced dataset for training.

Model Performance:

- We built and evaluated two models: Logistic Regression and Decision Tree.
- Logistic Regression showed strong performance with an ROC-AUC score of 0.9525. The classification report also indicated good precision, recall, and f1-score for both classes.
- Decision Tree also performed well, with an ROC-AUC score of 0.8377. While slightly lower than Logistic Regression, it still provided a decent level of accuracy in predicting churn.
- Feature importance from the Decision Tree model highlighted that PaymentMethod_Electronic check, tenure, and MonthlyCharges were among the most influential features in predicting churn.

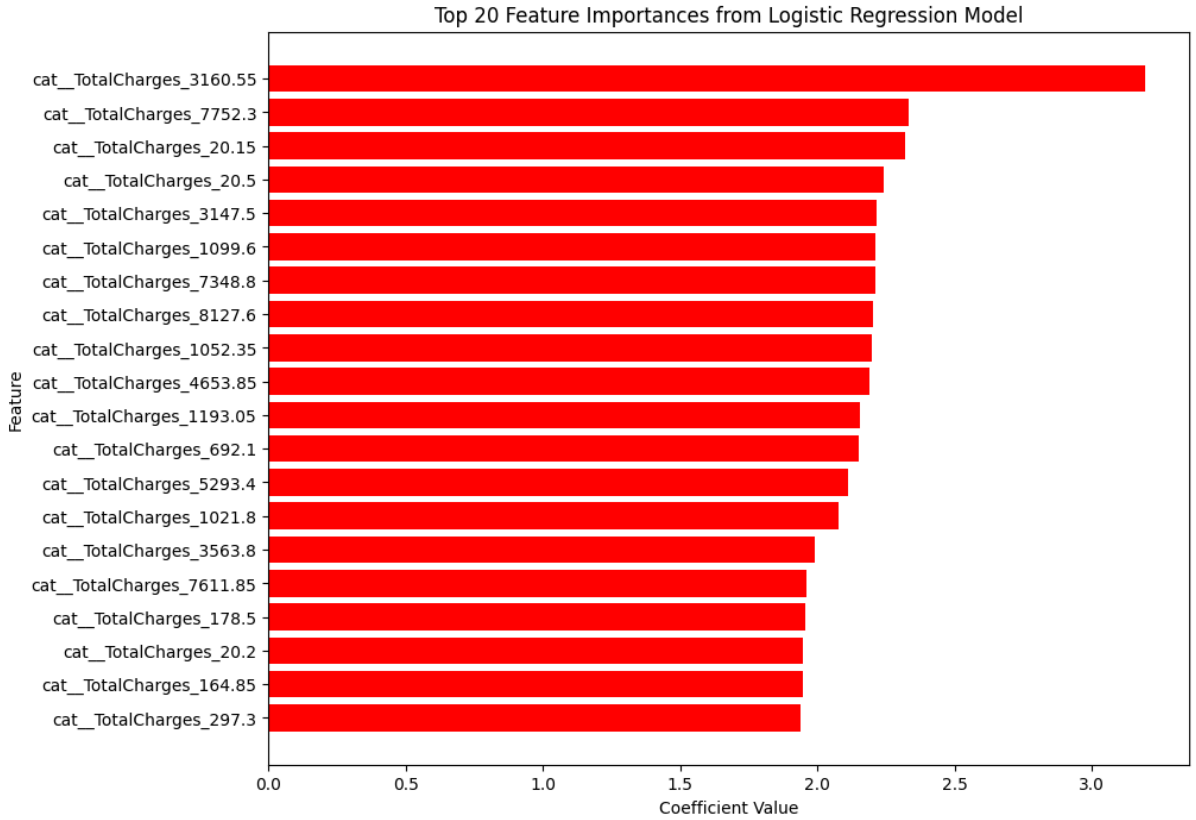
In summary, both models provide valuable insights into customer churn. The Logistic Regression model, in this case, appears to have slightly better predictive performance based on the ROC-AUC score. The feature importance from the Decision Tree gives us an idea of which factors are most strongly associated with churn according to that model.

Data Analysis Key Findings

- The dataset contains 7043 customer records and 21 columns. The target variable, 'Churn', has a class imbalance, with 5174 non-churners and 1869 churners.
- After handling categorical features and scaling numerical features, the training dataset size is 5634 records with 30 columns.
- Logistic Regression achieved an accuracy of approximately 80.3% and an F1-score of approximately 59.3% on the testing data.
- Decision Tree Classifier achieved an accuracy of approximately 79.1% and an F1-score of approximately 56.6% on the testing data.
- Neural Network Classifier achieved an accuracy of approximately 80.4% and an F1-score of approximately 60.2% on the testing data.
- Among the models evaluated, the Neural Network Classifier showed slightly better performance in terms of both accuracy and F1-score.
- Key features identified across models for predicting churn include 'Contract_Month-to-month', 'TotalCharges', 'InternetService_Fiber optic', and 'PaperlessBilling_Yes'.

Insights or Next Steps

- The class imbalance in the churn data is a significant factor impacting model performance, as indicated by the difference between accuracy and F1-scores. Further exploration of advanced techniques beyond SMOTE, such as exploring different oversampling/undersampling ratios or cost-sensitive learning, could potentially improve the F1-score and better capture churners.
- Focusing on customers with month-to-month contracts, higher total charges, fiber optic internet, and paperless billing could be a strategic area for targeted churn prevention efforts, given their high feature importance across models.



This bar chart highlights the top 20 most influential features from a logistic regression model. All features are binned values of `TotalCharges` , showing how different spending levels impact the prediction. The feature `cat__TotalCharges_3160.55` has the highest coefficient, indicating a strong positive influence on the model's output.