PHASE 3 PROJECT

CUSTOMER CHURN PREDICTION

OBJECTIVES

Main

1.Determine if existing customer data (e.g.demographics, call history, service plans)is relevant to predicting churn.

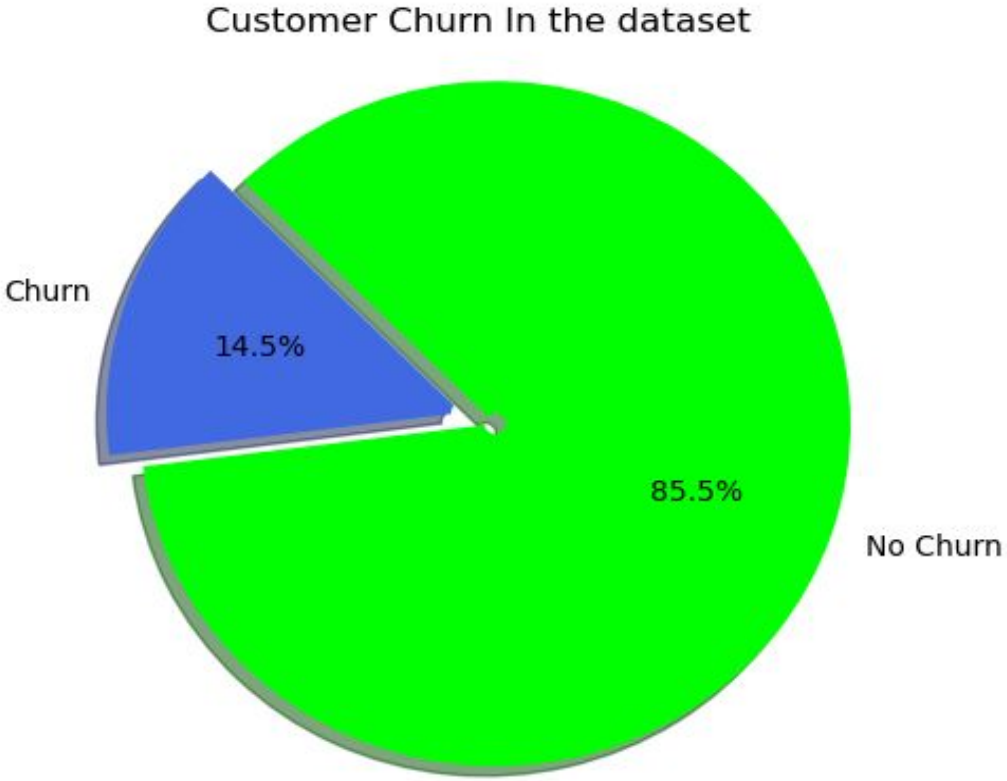
2.Evaluate the effectiveness of different retention and interventions in reducing churn rates .

3.Identify factors that lead to customer churn such as service quality,pricing and contract terms.

4. Segment customers based on their propensity to churn and tailor retention strategies accordingly.

TABLE OF CONTENT

CUSTOMER CHURN DISTRIBUTION



This pie chart shows the distribution of customer churn in the dataset. Approximately 14.5 % of customers have churned, while 85.5% have not churned. This indicates that the dataset is imbalance, with a higher proportion of non-churned customers compared to churned customers. We will need to consider this class imbalance when building and evaluating the churn prediction model.

Feature importance and selection.

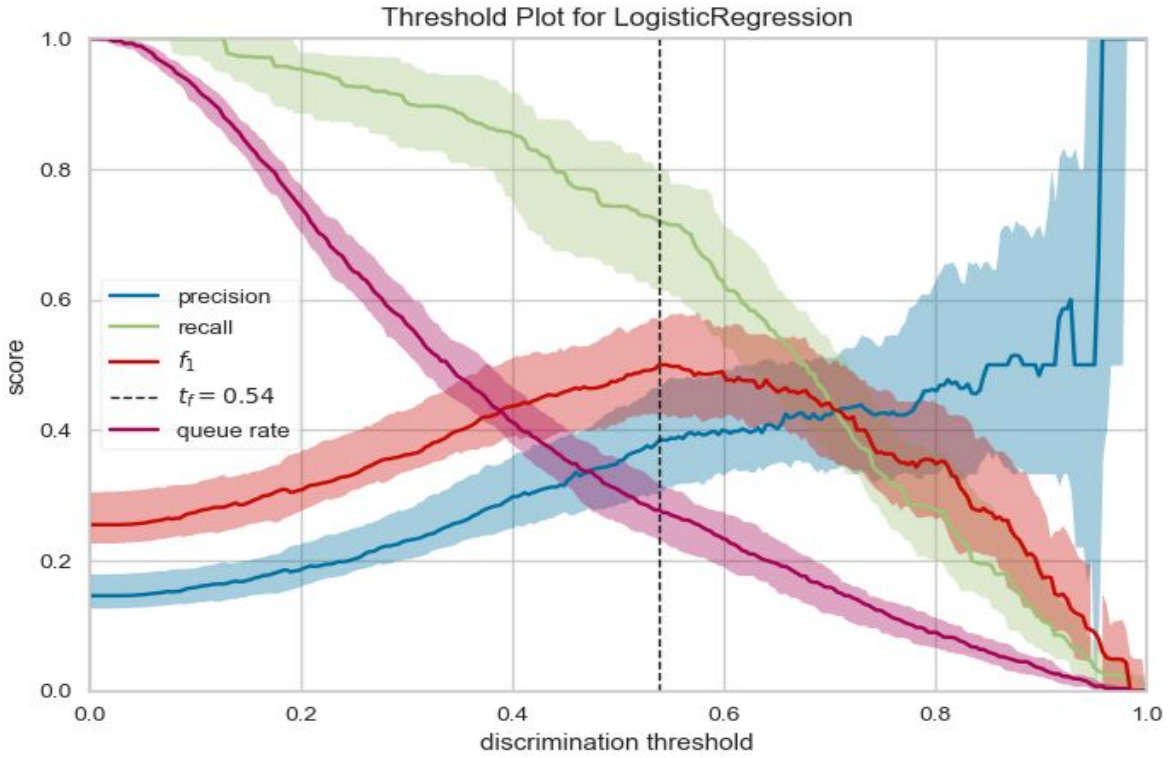
Anova Test - Based on the Anova test results, all numerical features are statistically significant and important for predicting customer churn. This suggests that all numerical features should be included in the churn prediction model.

Chi-square helps select relevant features especially for categorical data. It asses the relationship between a categorical feature and a target variable:

Null Hypothesis :No association between the feature and the target variable.

Alternative Hypothesis: The feature is relative to the target variable.

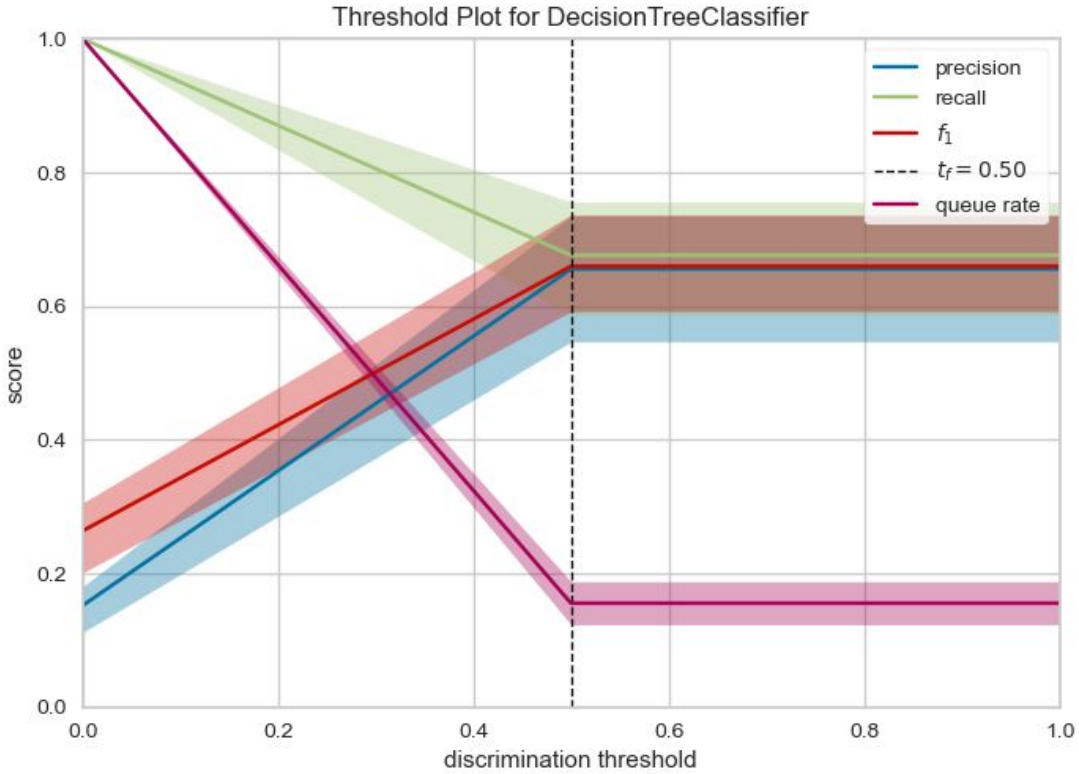
Model Building and Training - Logistic Regression



The Logistic Regression model used in this context, with balanced class weights, has shown a good ability to predict customer churn.

It has an accuracy of approximately 77%, and the ROC curve indicates a good balanced between sensitivity and recall.

Decision Tree Classifier



The Decision Tree model has shown a good ability to predict customer churn, with an accuracy of approximately 91%. The ROC curve indicates a good balance between sensitivity and recall.

ENSEMBLES

Random Forest Classifier

The RandomForestClassifier exhibits strong performance on the test data with a high accuracy of 94.0%, a balanced recall of 61.96% and precision of 91.94%, resulting in a commendable F1 score of 74.03%. On the training data, the model demonstrates even higher accuracy(95.76%), improved recall (71.1%), and a perfect precision score of 100%,yielding an impressive F1 score of 83.11%. T he out-of-bag score,a validation metric for metric for random forest, stands at 93.06%, further indicating the models robustness and effectiveness in generalizing to unseen data.

Gradient Boosting Classifier

Comparing the Gradient Boosting Classifier with the earlier Random Forest Classifier , the Gradient Boosting model exhibits slightly higher accuracy on the test data (94.0% vs 93.2%). Additionally, the gradient Boosting model shows improved recall (68.48% vs. 61.96%) and precision (85.14% vs. 83.05%)on the test test. On the training data, the gradient Boosting model continues to outperform with higher accuracy (96.62% vs. 95.76%), recall (77.24% vs. 71.10%),precision (99.67% vs. 97.77%), and F1 SCORE (87.03% VS.83.14%)

1. Random Forest Classifier
   - Test Data: Accuracy 93.20%, Recall 61.96%, Precision 83.05%
   - Train Data: Accuracy 95.76%, Recall 71.10%, Precision 97.77%

2. Gradient Boosting Classifier
   - Test Data: Accuracy 94.00%, Recall 68.48%, Precision 85.14%
   - Train Data: Accuracy 96.62%, Recall 77.24%, Precision 99.67%

3. XGBoost Classifier
   - Test Data: Accuracy 91.60%, Recall 41.30%, Precision 95.00%
   - Train Data: Accuracy 92.35%, Recall 47.83%, Precision 100.00%

Recommendations

The ensemble stacking model outperforms the individual models, demonstrating the highest accuracy, recall, precision, and F1 score on both test and training data.

- The stacking model is recommended for its improved overall performance in predicting the target variable.

Thankyou.