Customer Churn Prediction

Objectives

- Perform exploratory analysis and extract insights from the dataset.
- Build a model to predict which customers will churn and discuss why you choose a particular algorithm.
- Establish metrics to evaluate model performance.

Interpretation from EDA:

1. Each customer is identified through a unique phone number. There are 19 independent variables used to predict the target feature – customer churn. In this dataset, customer churn is defined as the number of people who stopped being customers.

2. Only around 483 out of 3333 customers in the dataset have churned. This means that we are dealing with an imbalanced classification problem. We must perform some feature engineering to create a balanced training dataset before building the predictive model.

3. In the box plot of total day charge and total eve charge by churn, we can see that Customers who churned have a higher median charge than customers who have not churned.

4. Customers who do not use voice mail plan churn more often than other users.

5. There is a positive relationship between total day minutes and total day charge, total eve minutes and total eve charge, total night minutes and total night charge, total intl minutes and total intl charge. It's obvious because the amount charged to a customer affects minutes spent by a customer on call.

6. State vw has high number of customers who have not churned. State IA has low number of customers who has churned. However NJ and TX are the two states with high number of customer churn.

7. The top three states of customer service calls are WV (159),NY (142),OR (135) respectively.

Model selection and Interpretation:

1. Best Model based on accuracy and F1 Score is XGBoost. We are going to use XGBoost model for customer churn prediction.

2. Accuracy: 92.09%

   This indicates the overall correctness of the model. In this case, 92.09% of the predictions made by the model are correct.

3. Precision: 70.62%

   Precision measures the accuracy of the positive predictions made by the model. A precision of 70.62% means that out of all instances predicted as positive, 70.62% were true positives, while the remaining 29.38% were false positives.

4. Recall (Sensitivity): 78.12%

   Recall measures the ability of the model to capture all the positive instances. In this case, 78.12% of the actual positive instances were correctly identified by the model, while 21.88% were missed (false negatives).

5. F1 Score: 74.18%

   The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, considering both false positives and false negatives. A higher F1 score indicates a better balance between precision and recall.

6. ROC AUC Score: 86.30%

   The Receiver Operating Characteristic Area Under the Curve (ROC AUC) measures the model's ability to distinguish between the two classes. An ROC AUC of 86.30% is relatively good, indicating a high true positive rate and a low false positive rate.

7. The model performs well in terms of accuracy, precision, recall, and F1 score.

8. The ROC AUC score suggests that the model is effective in distinguishing between the classes.

Conclusion:

Overall, the model shows good performance, hence I'll strongly recommend XGBOOST model to predict customer churning.