Churn Prediction Model Summary

Introduction:

Most of the telecommunication companies face challenges in keeping customers with them. So, it is very necessary to predict customer churn in order to implement effective retention strategies. In this analysis, we developed a Churn Prediction Model using machine learning algorithms. The aim was to develop a predictive model that could identify customers likely to churn which will help companies to apply retention methods.

Methods:

There are 7043 observations and 33 columns in the dataset.

Data Cleaning:

- Missing values found in the "Total Charges" and "Churn Reason" columns.
- All the missing values were handled using appropriate methods.
- "Total Charges" column was converted from Object to Numeric.

Exploratory Data Analysis:

- The company currently has a churn rate of 26.5% which indicates that a significant portion of customers have already stopped using its services.
- Approximately 33.23% of churned customers left because a competitor provided a more attractive offer.
- About 17.49% of churned customers left due to perceived bad behavior.
- It is necessary to monitor early-stage customers as they tend to abandon the service shortly after starting.
- Longer-term customers are more likely to stay.

Geographic Analysis:

- All customers are located in the United States, specifically in California.
- The data encompasses a total of 1129 cities within the state.
- A significant concentration of customers is observed in major cities such as Los Angeles, Long Beach, San Francisco, San Jose, and Berkeley.
- Los Angeles stands out as the city with the highest customer count with 305 customers.
- It is observed that a high portion of churn occurs in Los Angeles, possibly linked to its high customer density.

- There is a correlation between monthly charges and churn rates. Higher monthly charges are associated with higher churn rates, particularly when charges exceed $70 approximately.

Demographic Analysis:

- Both Male and Female non-senior citizens exhibit a total churn of 694 and 699, respectively, demonstrating a nearly similar distribution of churn.
- Similarly, senior citizens Males and Females show a total churn of 236 and 240, respectively.
- Churn has more occurrence among customers without partners (1200 churns) compared to those with partners (699 churns).
- Similarly, customers without dependents (1763 churns) experience higher churn than those with dependents (106 churns).
- For Female non-senior citizens, churn reasons include support personnel attitude, competitors offering more data, higher download speeds, and better overall offers.
- Non-senior citizens, both Male and Female, are more likely to churn.
- For Male non-senior citizens, churn reasons include support personnel attitude, competitors offering higher download speeds, more data, and unknown factors.
- Therefore, Churn is influenced by factors such as customer, competition, and product-related considerations.

Service Utilization Analysis:

- Customers with longer tenure have a higher likelihood of remaining with the company.
- The majority of customers, including churned ones, are connected to Fiber Optic Internet.
- 69.4% of churned customers were specifically using Fiber Optic Internet.
- Fiber Optic Internet Service users primarily churned due to the support person's attitude and the competitor offering higher download speeds. DSL service customers, in contrast, predominantly churned because the competitor provided higher download speeds and better devices.
- For Fiber Optic Internet users, subscribing to tech support significantly reduces the likelihood of churn.
- Customers using electronic checks for payment have the highest churn rate (45.3%).
- Identified three distinct groups based on monthly charges: Low Charge Group (less likely to churn), Medium Charge Group (moderately to churn), and High

Charge Group (most likely to churn). So, it is evident that monthly charges play a significant role in influencing customer churn.

Data Preprocessing:

- Categorical columns were encoded using Label Encoding.
- SMOTE (Synthetic Minority Over-sampling Technique) has been used to balance the dataset.
- The dataset was split into training and testing sets.

Model Building:

Logistic Regression:

- Achieved a training accuracy of 78.55% and a testing accuracy of 78.12%.
- The Cross-Validated Mean ROC AUC score was 86.56%.

Random Forest:

- Achieved a testing accuracy of 81.50% by tuning the Hyperparameter using GridSearchCV.
- The Cross-Validated Mean ROC AUC score was 93.89%.

Light Gradient Boosting (LGBM):

- Achieved a training accuracy of 90.19% and a testing accuracy of 86.68%.
- The Cross-Validated Mean ROC AUC score was 94.73%.

XGBoost Classifier:

- Achieved a training accuracy of 88.79% and a testing accuracy of 87.00%.
- The Cross-Validated Mean ROC AUC score was 94.47%.

Key Results:

- All models show strong ROC AUC Scores.
- XGBoost performed quite well than other models, achieving high accuracy score, precision, recall and F1-score.

Insights:

- The XGBoost model shows high performance across all the evaluation metrics.
- High ROC AUC indicates outstanding discrimination ability.
- Precision, recall, and F1-score are well-balanced, suggesting a strong model.
- The model performs well on the testing dataset, as shown by the almost similar training and testing accuracies.

Deployment Plan:

- Develop a web-based application named “Churn Prediction Model System”.
- Use Django framework for backend development and HTML, CSS, and JavaScript for frontend development. Also use appropriate database servers, security measures, and documentation.
- Continuously monitor and maintain the application, addressing any performance or security issues.

Conclusion:

In summary, the Churn Prediction Model System offers a powerful tool for telecommunication companies to enhance customer retention strategies and reduce churn rates.

NOTE: There is a thorough result explanation and deployment plan in the “Google Colab Notebook” File.