CHURN DATASET ANALYSIS

Machine learning

Name: G.Shashank

Email: gudimetlashashank999@gmail.com

Git-hub link : https://github.com/gudimetla-shashank/churn_analysis

Churn Prediction Project Report

Objective

The aim of this project is to predict customer churn for a telecommunications company using machine learning techniques. Customer churn, or attrition, occurs when customers stop doing business with a company. Accurate churn prediction allows companies to identify at-risk customers and take measures to retain them, thus reducing revenue loss.

Project Steps

1. Data Preprocessing

1. **Cleaning:** The dataset is cleaned to remove or handle missing values.
2. **Encoding:** The variable of object type are converted to the int or float type which are required.
3. After that checking for the null values after the encoding
4. Converting the unique types like “No internet service” to “No” for the required variables**.**
5. **Handling Outliers** : handling the outliers for the required variables using the IQR method.

2. Exploratory Data Analysis (EDA) Insights

During the EDA phase of this project, several key insights were obtained regarding the factors influencing customer churn. These insights are crucial for understanding customer behaviour and identifying at-risk customers. The following points summarize the findings:

1. **Demographic Distribution:**
   - The dataset shows an almost equal distribution of male and female customers.
2. **Internet Service Preferences:**
   - Among all available internet services, the majority of customers prefer Fiber optic services.
3. **Payment Methods:**
   - The most commonly used payment method is 'electronic check.'
4. **Contract Types:**
   - Approximately 50% of the customers opt for month-to-month contracts, rather than one-year or two-year contracts.
5. **Churn Analysis:**
   - **Internet Service:** Customers using fiber optic internet services are more likely to churn compared to those using other types of internet services.
   - **Payment Method:** Customers who use the electronic check payment method have a higher likelihood of churning.
   - **Contract Type:** Customers with month-to-month contracts exhibit a higher churn rate than those with longer-term contracts.
   - **Marital Status:** Customers without partners are churning at a higher rate compared to those with partners.
   - **Technical Support:** Lack of technical support is associated with higher churn rates.

- **Billing Method:** Customers who use paperless billing are more likely to churn compared to those who do not.
- **Monthly Charges:** The analysis shows that customers who churn have higher median monthly charges.

6. **Customer Independence:**
   - Independent customers, defined as those without partners, are less likely to churn.

3. Data preparation

Before diving into model building and evaluation, it is crucial to preprocess the data to ensure it is clean and suitable for analysis. One key aspect of data preprocessing involves encoding categorical variables into numerical values. This step is essential because most machine learning algorithms require numerical input.

1. **Label encoding** :

Label encoding was applied to convert categorical features into numerical values. The columns encoded were: `['gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'Churn']`.

2. **Train-Test split :**

The dataset was split into training and testing sets to evaluate the model's performance on unseen data. A 70-30 split was used, where 70% of the data was used for training and 30% for testing.

3. **Standard Scaling :**

Standard scaling was performed to normalize the feature values, ensuring that all features have a mean of 0 and a standard deviation of 1. This step is crucial for algorithms that are sensitive to the scale of the data.

4. **Feature Engineering using VIF**

Variance Inflation Factor (VIF) was used to identify and remove features with high multicollinearity. Features with high VIF values were removed to improve the model's performance and reduce redundancy.

5. **Balancing the Data using Random Oversampling**

To address the issue of class imbalance in the target variable, random oversampling was applied to the training data. This technique involves duplicating examples from the minority class to ensure that the classifier receives an equal number of instances from both classes, thus improving its ability to predict the minority class accurately.

4. Model selection

When selecting the best model for predicting customer churn, several factors need to be considered. This involves balancing model performance metrics, understanding the trade-offs between different models, and ensuring that the chosen model generalizes well to unseen data. Here are the steps and rationale behind selecting the best model:

1. Performance Metrics

Key performance metrics to consider include:

- **Accuracy:** The proportion of correct predictions out of all predictions made. While useful, it might not be sufficient for imbalanced datasets.
- **Precision and Recall:** Precision measures the accuracy of positive predictions, while recall measures the ability to identify all positive instances.
- **F1 Score:** The harmonic mean of precision and recall, providing a single metric that balances both concerns.
- **AUC-ROC:** Measures the model's ability to distinguish between classes, useful for binary classification tasks.

As i have applied some of the models like logistic regression regression,support vector machine,decision trees,random forest,naïve bayes,ada boost classifier and gradient boosting classifier.

As keeping all the above factors in the mind among all the models the logistic regression has work very well with 74% accuracy.

But if the data is unbalanced then I am getting an accuracy of 80% which is a good accuracy but also keeping the factors like precision ,recall and f1 score using the balanced data set is really an efficient idea.

As the other models like decision tree and random forest are occurring with overfitting so to overcome I have done the hyper parameter tuning. Even though applying the grid search I could not see the good results.

Conclusion:

From the above comments I would like to conclude that among all the models for me logistic regression has worked well, keeping all the evaluation metrics in the mind.

Problems faced:

- Problem faced whether to select the balanced data set or the unbalanced dataset in terms of the accuracy
- Problem faced when the model occurred with the overfitting and tried to sole that with hyper tuning parameter