**Telecom Customer Churn Prediction**

Predicting Customer Churn Using Machine Learning Techniques

**Introduction:**

The Telco Churn ML project is a data science initiative focused on predicting customer churn within the telecommunications industry. Data scientists and business analysts recognize the pivotal importance of customer churn prediction in driving retention and loyalty strategies.

By harnessing machine learning algorithms and techniques, this project aims to provide insightful predictions that can guide proactive retention efforts, ultimately reducing customer attrition rates.

**Data Set Information:**

The dataset provided is the "Telco Customer Churn" dataset, which contains information on customers and features related to their churn, with a target variable being the "Churn" column. Here is a summary of some key details from the dataset:

* The dataset consists of 7043 rows (customers) and 21 columns (features).
* Each row represents a customer, with columns containing various customer attributes described in the metadata.
* The "Churn" column denotes whether a customer has churned or not and is the target variable.
* Some of the attributes included in the dataset are Customer ID, Gender, SeniorCitizen status, Partner status, Dependents status, Tenure, PhoneService availability, MultipleLines availability, InternetService provider, and OnlineSecurity status.

Additionally, the dataset includes the following information:

* + The dataset has a breakdown of male and female customers, each representing 50%.
  + The distribution of different InternetService providers is as follows: 44% have Fiber optic, 34% have DSL, and 22% are labeled as 'Other'.
  + PhoneService availability summary: 50% do not have phone service, 29% have it, and 22% are labeled as 'Other'.

**Problem Statement:**

* Develop a classification model to predict customer churn based on the provided dataset.
* Identify key customer attributes and their impact on churn prediction.
* Evaluate and compare the performance of different machine learning algorithms for churn prediction.
* Provide actionable insights and recommendations for the company to reduce customer churn and improve customer retention strategies.

By addressing this problem, businesses can proactively identify potential churners and take appropriate actions to retain customers, thus impacting their overall customer satisfaction and revenue.

**Data Cleaning:**

1. Handling Missing Values:
   * It was noted that there were no missing values present in the dataset. This was crucial as missing data can impact the accuracy of the analysis and modelling process. Dealing with missing values is essential in ensuring the dataset's integrity.
2. Converting Data Types:
   * No specific details were provided on data type conversions in the content provided. However, converting data types is a common data cleaning task where numerical values are converted to the appropriate data types (e.g., converting strings to numeric values) for analysis and modelling.
3. Addressing Outliers:
   * The presence of outliers can skew the analysis results. While no specific mention of dealing with outliers was made in the available information, identifying and handling outliers could be a crucial step in ensuring the dataset's quality.
4. Removing Duplicates:
   * Eliminating duplicate records from the dataset is important to avoid redundant information that could affect the outcomes of data analysis. It was not explicitly mentioned if duplicate records were removed during the data cleaning process.
5. Standardizing or Normalizing Data:
   * Standardizing or normalizing numerical variables can help bring different scales of variables to a similar range, which is important for certain machine learning algorithms. It was not stated if such standardization or normalization was performed in the cleaning process.

**Types of Data Visualizations:**

1. Bar Charts:
   * Bar charts were used to visualize the distribution of categorical variables such as gender, senior citizen status, partner status, dependents status, phone service availability, multiple lines availability, internet service provider, online security status, and tenure ranges. Bar charts are effective for displaying frequency distributions of categorical data.
2. Percentage Distribution Tables:
   * Percentage distribution tables were employed to present the distribution of different customer attributes, such as gender distribution (50% male and 50% female), senior citizen status distribution, partner and dependents status distribution, tenure distribution, phone service availability, multiple lines availability, internet service provider distribution, and online security status distribution.

**Visualizations did on the dataset:**

1. Customer Gender Distribution:
   * The dataset includes a breakdown of male and female customers, with each gender representing 50% of the customer base.
2. Senior Citizen Status Distribution:
   * The dataset indicates the percentage of customers who are senior citizens and those who are not. The breakdown shows that 58% of customers are not senior citizens, 42% are senior citizens, and 10% are labeled as 'Other'.
3. Partner and Dependents Status Distribution:
   * The dataset contains information about whether the customer has a partner or dependents.
4. Tenure Distribution:
   * The distribution of customer tenure in months was provided, showing the count of customers in different tenure ranges.
5. Phone Service and Multiple Lines Availability Distribution:
   * The availability of phone service and multiple lines for customers was summarized, indicating the percentage of customers with and without these services.
6. Internet Service Provider Distribution:
   * The dataset includes information about the distribution of different internet service providers among the customers, with 44% having Fiber optic, 34% having DSL, and 22% labeled as 'Other'.
7. Online Security Status Distribution:
   * The availability of online security for customers was presented, specifying whether customers have online security, no online security, or no internet service.

**Result:**

* High Churn seen in case of Month to month contracts, No online security, No Tech support, First year of subscription and Fibre Optics Internet.
* LOW Churn is seen in case of Long term contracts, Subscriptions without internet service and The customers engaged for 5+ years.
* Higher Monthly Charge at lower tenure results into lower Total Charge. Hence, Higher Monthly Charge, Lower tenure and Lower Total Charge are linked to High Churn

**Pre-processing:**

* Created a copy of the original DataFrame: df\_copy = df.copy()
* Dropped the 'tenure' column as data was already extracted from it: df\_copy.drop('tenure', axis=1, inplace=True)

**Encoding:**

* Selected categorical columns from df\_copy excluding 'Churn':

****

* Applied Label Encoding to transform categorical columns into numerical values:

A screenshot of a computer

Description automatically generated

* Replaced 'Churn' column values 'No' with 0 and 'Yes' with 1:

**Chi-square test**

* The Chi-Square test is a statistical method used to determine whether there is a significant association between two categorical variables.
* It tests whether the observed frequencies of categories differ significantly from the frequencies expected under independence.
* A p-value is obtained from the test, where a low p-value indicates statistical significance and rejects the null hypothesis of independence.
* It is commonly applied in research and data analysis to assess relationships between categorical variables and determine if observed differences are statistically significant.
* We utilized the Chi-Square test to analyze relationships between categorical variables in the dataset.
* It’s been used to test for significant associations between customer churn (binary) and other categorical variables like customer demographics or services subscribed.

1. **Feature Transformation:**
   * Encoding Categorical Variables: The categorical variables were encoded using Label Encoding to convert them into numerical values, making them suitable for machine learning algorithms.
2. **Data Splitting:**
   * Train-Test Split: The dataset was likely split into training and testing sets to evaluate the model's performance effectively.
   * X and y Definition: Features (X) and the target variable (y) were defined for model training and evaluation.
3. **Oversampling:**
   * Addressing Class Imbalance: Oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), may have been applied to balance the class distribution of the target variable, especially in cases of imbalanced data where one class dominates the other.
   * Oversampling Minority Class: Synthetic samples were likely generated for the minority class to increase its representation in the dataset and improve model performance on predicting the minority class.

**Model Building:**  
**Logistic Regression:**

**Model Building:**

* Instantiated a logistic regression model (lr=LogisticRegression()) for predicting telecom customer churn.
* Trained the model using the dataset.

**Performance Evaluation:**

* Achieved a training accuracy of 82.51% and a testing accuracy of 77.12%.

**Classification Report:**

**Precision:**

* + For predicting "No churn" (customer retention), achieved 87% precision, and for "Churn" (customer leaving), achieved 56% precision.

**Recall:**

* + Identified 81% of actual "No churn" cases and 66% of actual "Churn" cases

**F1-score:**

* + Harmonic mean of precision and recall was 84% for "No churn" and 60% for "Churn".

**Decision Tree:**

**Model Building:**

* + Instantiated a decision tree classifier and trained it using the dataset.

**Performance Evaluation:**

* + Achieved an exceptionally high training accuracy of 99.82% and a testing accuracy of 70.87%.

**Classification Report:**

**Precision:**

* + - Precision for predicting "No churn" was 83%, and for "Churn" was 46%.

**Recall:**

* + - Recall for identifying "No churn" instances was 76%, and for "Churn" instances was 57%.

**F1-score:**

* + - F1-score, the harmonic mean of precision and recall, was 79% for "No churn" and 51% for "Churn”.

**Random Forest Classifier:**

**Model Building:**

* + Employed a random forest classifier and trained it using the dataset.

**Performance Evaluation:**

* + Achieved a high training accuracy of 99.82% and a testing accuracy of 78.19%.

**Classification Report:**

**Precision:**

* + - Precision for predicting "No churn" was 86%, and for "Churn" was 58%.

**Recall:**

* + - Recall for identifying "No churn" instances was 84%, and for "Churn" instances was 63%.

**F1-score:**

* + - F1-score, the harmonic mean of precision and recall, was 85% for "No churn" and 60% for "Churn".

**Adaboost Classifier:**

**Model Building:**

* + Utilized AdaBoost with 50 estimators and a learning rate of 1.0, using the SAMME.R algorithm, and trained it using the dataset**.**

**Performance Evaluation:**

* + Achieved a training accuracy of 81.73% and a testing accuracy of 75.81%.

**Classification Report:**

* + **Precision:**
    - Precision for predicting "No churn" was 89%, and for "Churn" was 53%.
  + **Recall:**
    - Recall for identifying "No churn" instances was 77%, and for "Churn" instances was 73%.
  + **F1-score:**
    - F1-score, the harmonic mean of precision and recall, was 82% for "No churn" and 62% for "Churn".

**Gradient Boosting Classifier:**

* **Model Building:**
  + Utilized Gradient Boosting with a learning rate of 0.1 and 100 estimators and trained it using the dataset.
* **Performance Evaluation:**
  + Achieved a training accuracy of 83.83% and a testing accuracy of 77.06%.
* **Classification Report:**
  + **Precision:**
    - Precision for predicting "No churn" was 88%, and for "Churn" \was 55%.
  + **Recall:**
    - Recall for identifying "No churn" instances was 79%, and for "Churn" instances was 70%.
  + **F1-score:**
    - F1-score, the harmonic mean of precision and recall, was 84% for "No churn" and 62% for "Churn".

**Performances:**

* A graph of blue and orange bars

  Description automatically generated**Logistic Regression:** Achieved a training accuracy of 82.51% and a testing accuracy of 77.12%. Demonstrated balanced performance in predicting both "No churn" and "Churn" cases.
* **Decision Tree Classifier:** Showed overfitting with a high training accuracy of 99.82% but a lower testing accuracy of 70.87%.
* **Random Forest Classifier:** Achieved high accuracy on both training (99.82%) and testing (78.19%) datasets, indicating robustness and generalization capability.
* **AdaBoost Classifier:** Demonstrated moderate accuracy on both training (81.73%) and testing (75.81%) datasets. Showed good performance in predicting "No churn" cases but relatively lower performance in predicting "Churn" cases.
* **Gradient Boosting Classifier:** Showed decent accuracy with a training accuracy of 83.83% and a testing accuracy of 77.06%. Similar to AdaBoost, it exhibited good performance in predicting "No churn" cases but lower performance in predicting "Churn" cases compared to Random Forest.

**Conclusion**

Our evaluation of machine learning algorithms highlights the Random Forest Classifier as the top performer, offering high accuracy and balanced predictive capabilities. Random Forest stands out for its robustness and suitability for real-world deployment. Implementing this model can empower telecom companies to proactively retain customers and drive business growth.

Result:

High Churn seen in case of Month-to-month contracts, No online security, No Tech support, First year of subscription and Fibre Optics Internet.

Low Churn is seen in case of Long-term contracts, Subscriptions without internet service and The customers engaged for 5+ years.