**CLASSIFICATION PROJECT**

Predicting Customer Churn

**Introduction and Background**

Classification is a type of supervised machine learning where the goal is categorizing data into predefined classes or labels based on input features. Using labeled data models are trained to predict future occurrences, which is important to business executives and management.

This project is on predicting customer churn in the telecom industry. The churn rate is a metric used to measure the rate at which customers leave or cancel their service subscriptions. This impacts companies’ revenue, brand reputation, and market share. It is therefore important to find ways of reducing churn through proactive measures for sustainable growth and profitability in a competitive industry.

**Data Collection and Preprocessing**

The dataset used for this project was provided by my instructors and was extracted from three sources. The first dataset was extracted from Microsoft SQL Server, the second dataset was downloaded from a OneDrive, and the third dataset was downloaded from a Github repository.

The structure of the dataset is as follows:

Each row represents a respondent to the survey. There are roughly 5000 participants to the train dataset and the remaining which is the third dataset (test dataset). Below are the details on the columns for data understanding;

* CustomerID => Is the unique identification for each client
* Gender => Whether the customer is a male or a female.
* SeniorCitizen => Whether a customer is a senior citizen or not.
* Partner => Whether the customer has a partner or not (Yes, No).
* Dependents => Whether the customer has dependents or not (Yes, No)).
* Tenure => Number of months the customer has stayed with the company
* PhoneService => Whether the customer has a phone service or not (Yes, No).
* MultipleLines => Whether the customer has multiple lines or not.
* InternetService => Customer's internet service provider (DSL, Fiber Optic, No).
* OnlineSecurity => Whether the customer has online security or not (Yes, No, No Internet)
* OnlineBackup => Whether the customer has online backup or not (Yes, No, No Internet)
* DeviceProtection => Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport => Whether the customer has tech support or not (Yes, No, No internet)
* StreamingTV => Whether the customer has streaming TV or not (Yes, No, No internet service)
* StreamingMovies = > Whether the customer has streaming movies or not (Yes, No, No Internet service).
* **Contract**=> The contract term of the customer (Month-to-Month, One year, Two year)
* PaperlessBilling => Whether the customer has paperless billing or not (Yes, No)
* Payment Method => The customer's payment method (Electronic check, mailed check, Bank transfer(automatic), Credit card(automatic))
* MonthlyCharges => The amount charged to the customer monthly.
* TotalCharges => The total amount charged to the customer.
* Churn => (target variable) Whether the customer churned or not (Yes or No).

**Exploratory Data Analysis (EDA)**

Python was used throughout this project and it began with first installing packages and libraries needed. Some of the packages were pandas, numpy, matplotlib, and scikit-learn. The packages needed for this project to run are listed in the requirements.txt file in my repository listed.

The dataset was then loaded and preprocessed. In the data preprocessing phase, we check the columns to reflect what was listed in the data understanding. We also study the data to understand the data before beginning work. We further check for null values, impute based on our understanding of the dataset, check for duplicates, and perform data cleaning for consistency. Since our test dataset was already separated, our task was to merge two datasets and use them for training. During inspection the following were observed:

* The data has the same columns and column names.
* There were different datatypes in some columns. For instance, SeniorCitizen, Partner, Dependents, PhoneService, PaperlessBilling, and TotalCharges had conflicting datatypes which was inconsistent with our data understanding.
* There were missing values detected in some columns.

Hence referring to the documentation on the dataset, we set out to clean our dataset. The plan was to clean them side-by-side before merging them into a training dataset for training our model. In all the following were done to get a clean and consistent training dataset:

* Numerical dtypes were imputed with the various column median. This was because there was skewness in the data which make median better option compared to the mean.
* Categorical dtypes with missing values were filled with ‘most frequent’ data.
* Columns which had differing dtypes were converted as indicated in the data understanding documentation.
* The dataset was then merge using the code below and saved accordingly

df=pd.concat([data1,data2])

df.to\_csv('train.csv', index=False)

Next, we perform a Univariate, bivariate, and multivariate analysis of our combined train dataset. Firstly our univariate analysis it to help us understand individual variables like the distribution, characteristics, and other properties of the variables. Bivariate is also to explore the relationship between variables, test our hypothesis and answer other research questions.

Fig 1: Distribution of Churn

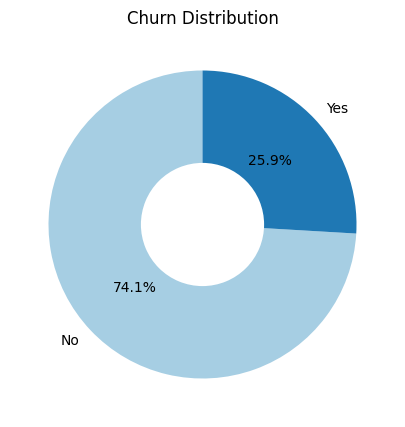


Fig 1 shows the distribution of Churn in our dataset, 25.9 percent churned while 74.1 percent did not churn. This indicates an imbalance which must be noted during modeling to prevent bias.

Fig 2: Distribution of Total Charges

A blue and white graph

Description automatically generated

Total charges as displayed in fig 2 shows a positively skewed distribution.

Fig. 3: Distribution of Payment Method

A chart with different colored squares

Description automatically generated

From the above Fig. 3 we see that most people pay using electronic check, the other means of payment Bank transfer, credit card, and mailed check followed respectively.

Research Questions

1. How does gender impact customer churn? Are there significant differences in churn rates between male and female customers?

Fig. 4: Churn Rate by Gender

A blue rectangular bars with white text

Description automatically generated

The churn rate as shown in fig. 4 indicates that there was little difference in churn rate by gender. 25.9 percent males churned while females who churned where 26.1 percent.

1. Does the presence of a partner or dependents influence customer churn? Is there a relationship between marital status and churn behavior?

Fig. 5: Churn rate by partner and by dependents

A comparison of blue and white bars

Description automatically generated

The presence of a partner and dependents have similar effects. Those with partners churn less same as those with dependents. It is not clear why this is the case, probably because individuals with partners and/or dependents had other reasons which must be investigated further.

1. How does the length of tenure affect churn rates? Are customers with longer tenures less likely to churn compared to new customers?

The fig. 6 shows that individuals with short tenure churned more.

Fig. 6: Churn rate by Tenure Group

A graph with numbers and lines

Description automatically generated

1. What role do additional services (eg. Online security, tech support, streaming TV, etc) play in reducing customer churn? Are customers with more services less likely to churn?

Fig. 7: subplots of churn to different services provided.

Most people with additional services churned less compared to people without these services

1. Is there a correlation between the contract term and churn rates? Are customers on long-term contracts less likely to churn compared to those on month-to-month contracts?

Fig. 7: Types of contract vs churn rate

A graph of different colored bars

Description automatically generated

As shown in fig. 7 above, individuals with longer contracts churned less compared to shorter contracts. Is this because these short-term contract customers were attracted by short term promotional products and leave when these activities end? Or they were attracted by other similar competitive offers?

**Hypothesis testing with chi-square test.**

H\_o: Contract type does not influence customer churn

H\_a: Contract type influences customer churn

We employed chi-square test because we are testing the association of two categorical variables. This test is useful when investigating if there is a relationship between two variables that do not involve measurement on a continuous scale, but distinct categories or groups.

**Conclusion**

From our results we reject the null hypothesis since our p-vale is below our significance value of 5%. Therefore, contract type influences customer churn significantly. 89683084478856e-191

Degrees of Freedom: 2

**Machine Learning Model Building**

Data balancing

In building our models, we start by balancing our data since we noticed earlier that there was some imbalance. Heavily skewed classes have several effects on machine learning models, these include a bias towards the majority class, poor generalization, and inflated accuracy.

We use ‘RandomOverSampler’ to balance our data using a sampling strategy of 50 percent.

Data splitting

We then split our data into train and evaluation dataset using the train\_test\_split by 80-20 percent. This is to enable us evaluate our model before the final evaluation with the test dataset.

Numerical columns and categorical columns are separated and assigned to num\_imputer and cat\_imputer to use a simple imputer. These are later merged before we encode features using the OneHotEncoder. This encoding process converts categorical data into numerical format for compatibility with machine learning algorithms. We use the One-Hot encoding since our categories have no ordinal relationship.

New we do a features scaling to transform values of variables into common scale or range. This is done to ensure that different features have a similar influence on machine learning algorithms.

**Model Evaluation**

For our purposes we select and train five(5) models. These models include

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* Ada Boost Classifier
* Gradient Boosting Classifier

The following are the results of our training and evaluation test

|  |  |  |
| --- | --- | --- |
|  | Model | Accuracy Score |
| 1 | Logistic Regression | 75.14 |
| 2 | Decision Tree Classifier | 78.02 |
| 3 | Random forest Classifier | 82.43 |
| 4 | Ada Boost Classifier | 75.95 |
| 5 | Gradient Boosting Classifier | 76.94 |

We went further to do a cross-validation, which instead of a single train-validation-test split, we do a cross-validation. This involves repeatedly splitting the training data into different subsets(folds). This way we can perform multiple rounds of training and evaluation to get more reliable performance estimates.

In the cross-validation we still had Random Forest classifier with the highest mean accuracy of 82.74 percent, which is a slight improvement.

**Conclusion**

In summary, random forest performed best and is the best option to predict the churn rate of our dataset.