# Predicting Customer Churn in a Telecommunications Company

*Dissertation submitted in fulfilment of the requirements for the Degree of*

**BACHELOR OF TECHNOLOGY**

**in**

## COMPUTER SCIENCE AND ENGINEERING

By

## SAMRIDDHI GUPTA

**12107731**

**Submitted to:**

**SpeakX**



**School of Computer Science and Engineering**

Lovely Professional University Phagwara, Punjab (India)

June 2024

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| S.NO | TITLE | PG.NO |
| 1. | **INTRODUCTION** | **3** |
| 2. | **ABOUT THE DATASET** | **3** |
| 3 | **DATA PREPROCESSING** | **3** |
| 4. | **EDA** | **4** |
| 5. | **MODELS USED AND RESULT** | **6** |
| 6. | **CONCLUSION** | **8** |

**INTRODUCTION**

In the telecom sector, predicting customer attrition is crucial for businesses to keep existing clients and cut down on the expenses of bringing on new ones. Telecom firms must recognize and address the causes that lead to customer churn because the cost of keeping an existing client is far lower than the cost of obtaining a new one. The recent years have seen a significant increase in the application of machine learning models to forecast customer attrition based on a variety of variables, including demographics, payment history, and usage habits.

The objective of this project is to create a machine learning model that can reliably forecast customer attrition in the telecom sector. Our objective is to create a model capable of precisely identifying consumers who are most likely to go and offering advice to telecom providers on how to keep them around.

Data preprocessing, feature engineering, model selection, and hyperparameter tuning are some of the tasks that the project will entail. To forecast client attrition, we will employ well-liked machine learning methods including Random Forest, Support Vector Classifier, and Logistic Regression. The models' performance will be assessed through the use of metrics including F1-score, recall, accuracy, and precision.

**ABOUT THE DATASET**

The dataset for this project is taken from Kaggle. We will train our ML model using the Labeled data set this type of training we call it as Supervised Machine Learning.

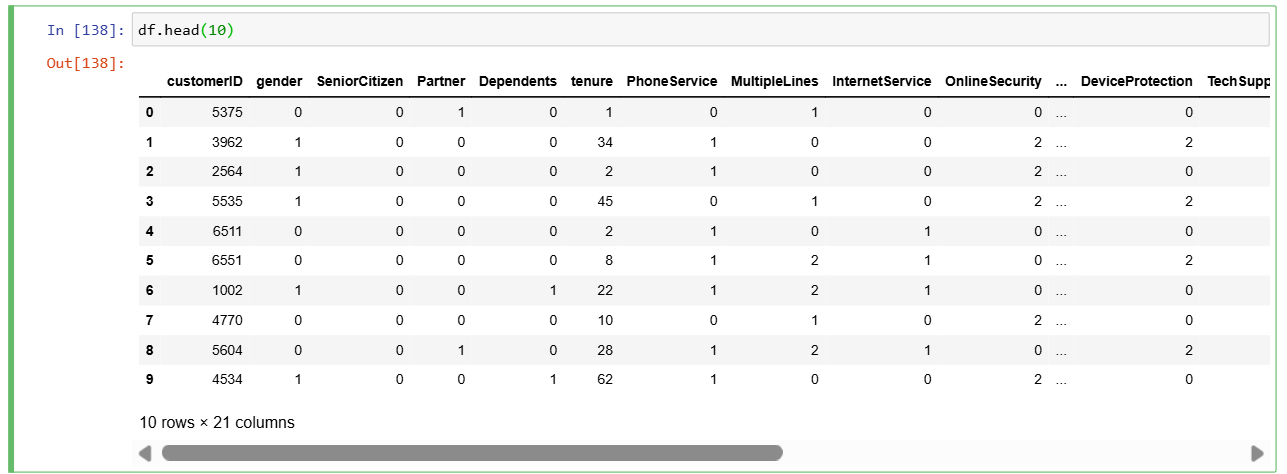
The Telco Customer Churn dataset is a comprehensive collection of customer data from a telecommunications company. The dataset offers a wealth of insights for comprehending consumer behaviour and forecasting attrition because it includes data on customer demographics, usage trends, and service preferences. The dataset provides a comprehensive picture of customer attributes, such gender, partner status, number of dependents, phone service usage, and internet service type, with 7,043 rows and 21 columns. The collection also contains details online security, backup, streaming TV and movies, paperless billing, payment methods, and contract types. This information can be used to create predictive models that can identify at-risk consumers and guide focused retention initiatives because it includes the "Churn" variable, which indicates whether a client has left the service provider.

**DATA PREPROCESSING**

**One hot Encoding**

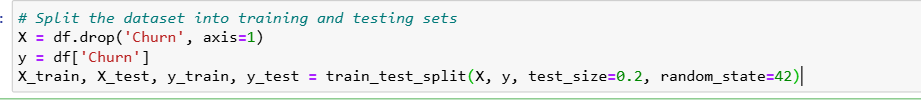
I used the LabelEncoder from scikit-learn to encode the categorical variables in my dataset. This is necessary because many machine learning algorithms cannot handle categorical variables directly. The LabelEncoder converts the categorical variables into numerical format, making them suitable for use in machine learning models.

I applied the LabelEncoder to the following categorical variables: gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, TechSupport, OnlineSecurity, OnlineBackup, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod, Contract, Churn, customerID, DeviceProtection, MonthlyCharges, and TotalCharges. This process ensures that the categorical variables are properly encoded and can be used as input features for my machine learning model.

****

**Data Splitting**

The train\_test\_split tool from scikit-learn was used to divide the dataset into training and testing sets. This is an important machine learning stage since it lets us assess how well our model works with unknown data. The training set and the testing set are the two halves of the dataset. 80% of the data made up the training set, which was used to train the model, and 20% of the data made up the testing set, which was used to assess the model's performance.



**EXPLORATORY DATA ANALYSIS**

Exploratory data analysis(EDA) is a crucial step in data science, which entails analyzing, examining, and visualizing data to extract significant insights. Comprehending the fundamental organisation of a dataset, recognising patterns, correlations and trends among variables, and developing conjectures for additional research are beneficial.

**Checking Missing Values**

We checked for thr missing values and the outcome confirms that the dataset has no missing values, which is crucial for guaranteeing the dependability and quality of my analysis.

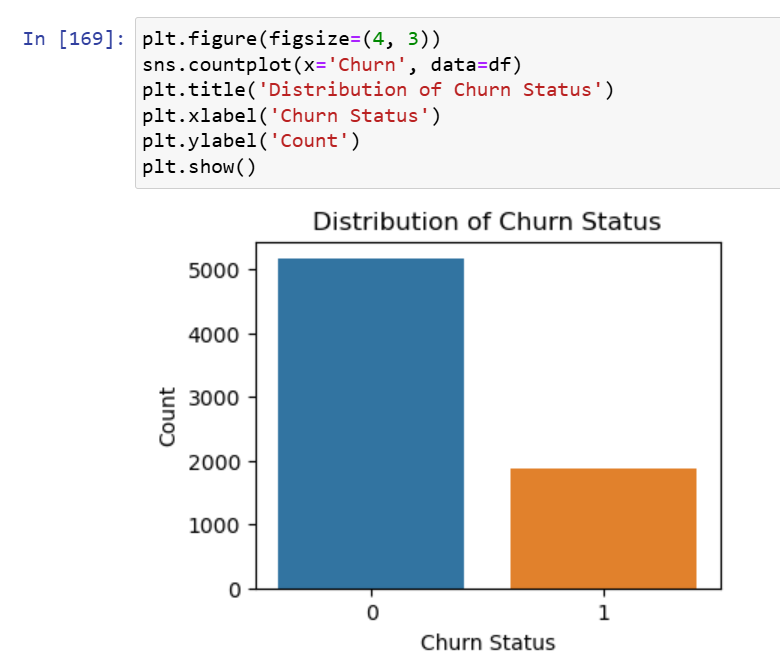
This check is essential because missing values can negatively impact my model's performance and result in false results.

A screenshot of a computer program

Description automatically generated

**Count Plot**

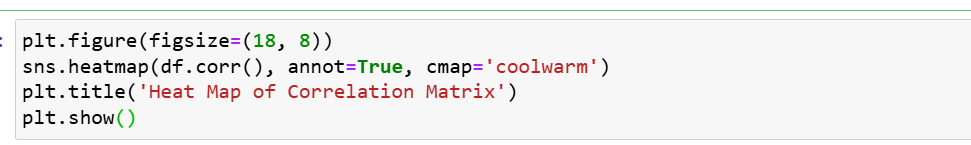
This plot gives me important insights into the issue of client retention by showing me how many customers have churned and how many have not. With a much higher proportion of non-churned customers than churned customers, the count plot amply demonstrates the dataset's imbalance. I created a count plot to visualize the distribution of churn status in my dataset using the following code:

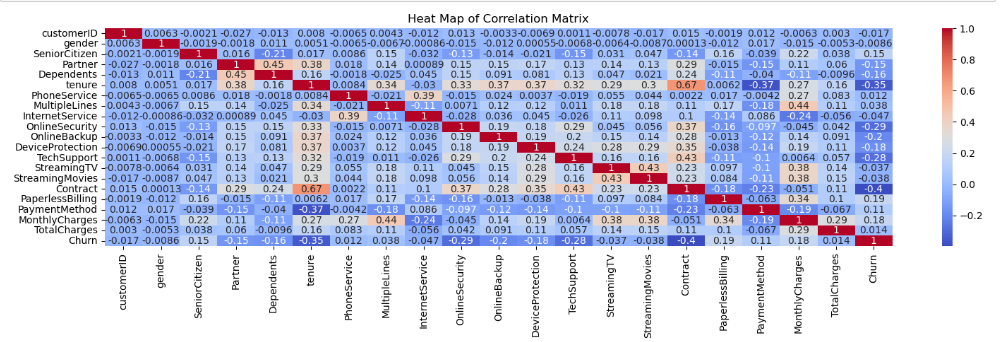


**Heat map of Coefficient Matrix**

The correlation between the various variables in my dataset is shown graphically in this heatmap in an understandable manner. The degree of the link is shown by the colour intensity in the heatmap; greater correlations are indicated by darker colours. By examining the heatmap, I can identify which variables are highly correlated with each other.

The heatmap also allows me to quickly identify which variables are not correlated with each other, which can be useful for selecting features that are not redundant or highly correlated.





**MODELS USED AND RESULTS**

**The Random Forest Classifier**

Several decision trees are combined in the Random Forest Classifier, a well-liked ensemble learning technique, to increase the model's accuracy and resilience. It excels at managing high-dimensional data and is renowned for its capacity to manage outliers and missing values.

A screenshot of a computer program

Description automatically generated

**K-Nearest Neighbours Classifier**

A straightforward yet powerful classification technique, the K-Nearest Neighbours Classifier finds the training set's most comparable samples and uses their majority vote to forecast the class of a new sample. Since it is non-parametric, no presumptions regarding the data's distribution are necessary.

A screenshot of a computer program

Description automatically generated

**Support Vector Classifier**

For classification problems, a well-liked machine learning technique is the Support Vector Classifier (SVC). It's a kind of supervised learning algorithm that works on classification issues that are linear or non-linear. The SVC algorithm locates the hyperplane that maximises the distance between the nearest data points and the decision boundary, or the margin between the classes.

A screenshot of a computer program

Description automatically generated

**Logistic regression**

A well-liked classification approach called logistic regression models, using a logistic function, the likelihood that a sample will belong to a specific class. It is a commonly used, easily understood approach with a wide range of applications.

A screenshot of a computer program

Description automatically generated

**CONCLUSION**

In summary, the goal of this research was to use a range of machine learning models to forecast client attrition. The Telco Customer Churn dataset, which includes details on customer demographics, consumption trends, and service preferences, was the dataset that was used. The Random Forest Classifier, K-Nearest Neighbours Classifier, Support Vector Classifier, and Logistic Regression were the models that were employed.

With an accuracy of 0.81 and a classification report indicating high precision and recall for both positive and negative classes, the Logistic Regression proved to be the most effective, according to the data. With respective accuracies of 0.74 and 0.79, the K-Nearest Neighbours Classifier and Random Forest Classifier demonstrated strong performance. With an accuracy of 0.73, the Support Vector Classifier model had the lowest performance.

The research shows how well machine learning models predict customer attrition and emphasises how crucial it is to select the appropriate model for the given situation. Because of its high accuracy and resilience to outliers, the Random Forest Classifier is an excellent option for this problem, according to the results.