# Data analysis and machine learning with Python for real-world

# Introduction

## 1. 1 Background Overview

Modern organizations and companies collect massive amounts of information in the course of conducting business. If this data is examined properly, then it can be very useful for modifying future decision making procedures. To this extent, predictive analysis and machine learning are some of the power tools that organizations use to analyze such data in order to be in a position to predict future trends and/or behaviors.

## 1. 2 Objectives

Thus, the main goal of this paper is to highlight the areas of machine learning in business and its application to the case of customers’ churn prediction. More precisely, this report seeks to understand the potential of predictive analytics and analyze the efficiency of different types of machine learning algorithms.

## 1. 3 Scope of Study

In this research, customer churn prediction in the telecommunications industry is of interest and data from the Telco Customer Churn dataset are used. Data cleaning, exploratory data analysis, training of the model and evaluating the same will be part of this analysis.

## 1. 4 Role of the Predictive Analytics and Machine Learning for Data Driven Decision-making

Today business decisions cannot be made without predictive analytics and machine learning. Being able to learn from historical records, these technologies are capable of detecting patterns and thus predicting future occurrences, giving business the ability to act and prevent instead of acting on a projected aftermath. They also assist in enhancing the quality and speed of work such as customer loyalty, sales prediction, and risk control.

## 1. 5 Real world business and social problems solved by machine learning

The usage of machine learning is common in different fields. In business, it assists in achieving efficiency in the delivery of services to customers, identify fraud and cut on the expenses required to run the business. It has its use in health care where it is used in forecasting disease incidences and in the diagnosis of the disease. In societal pertinent issues it supports decision making on resources sharing, prediction of criminal activities, and disease control. It is also a rather universal option capable of providing reasonable approaches to the vast variety of issues facing businesses and societies.

# Exploratory Data Analysis

## 2.1 Dataset Selection and Problem Identification

Subset for this project is the Telco Customer Churn dataset collected from [Kaggle](https://www.kaggle.com/datasets/blastchar/telco-customer-churn). The first and focal business issue that has been addressed is to forecast the ‘churn’, meaning the customers who are likely to leave the telecommunication service. Since the program can accurately predict which customers are likely to defect, it can assist the telecom companies to design ways of protecting their customer base and therefore, contain the possible revenue loss.

## 2. 2 Data Types and the Structure

The dataset consists of the following data types:

* Categorical Variables: Gender, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Churn
* Numerical Variables: SeniorCitizen, tenure, MonthlyCharges, TotalCharges
* Target Variable: Churn (whether the customer walked away or not).

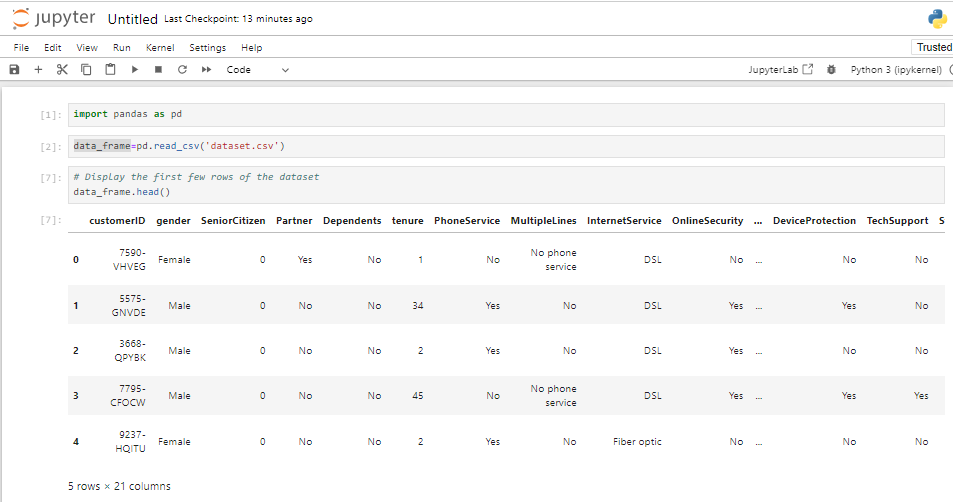


Figure 1: Jupyter Notebook: Data Loading and Initial Inspection

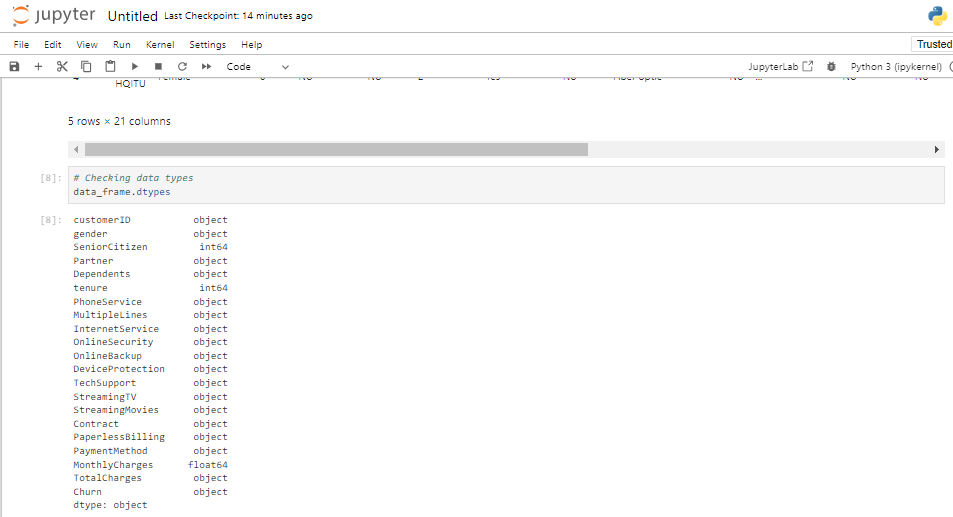


Figure 2: Jupyter Notebook Code and Output: Data Types Exploration

The used data set is composed of different customer characteristics including the taken services, explicit and implicit profile and account data. Thus, the aim is to maximize the accuracy in estimating the Churn column, which is binary in nature.

## 2.3 Data Cleaning and Encoding

It is important to preprocess the data before constructing the models of machine learning. This includes feature preprocessing to handle missing values, and also encoding of categorical variables for compatibility with the models.

* **Handling Missing Values**: In the dataset there may be some columns with missing or incorrect data, for example, the TotalCharges column. These will either be filled or dispensed, which are available in the table below;

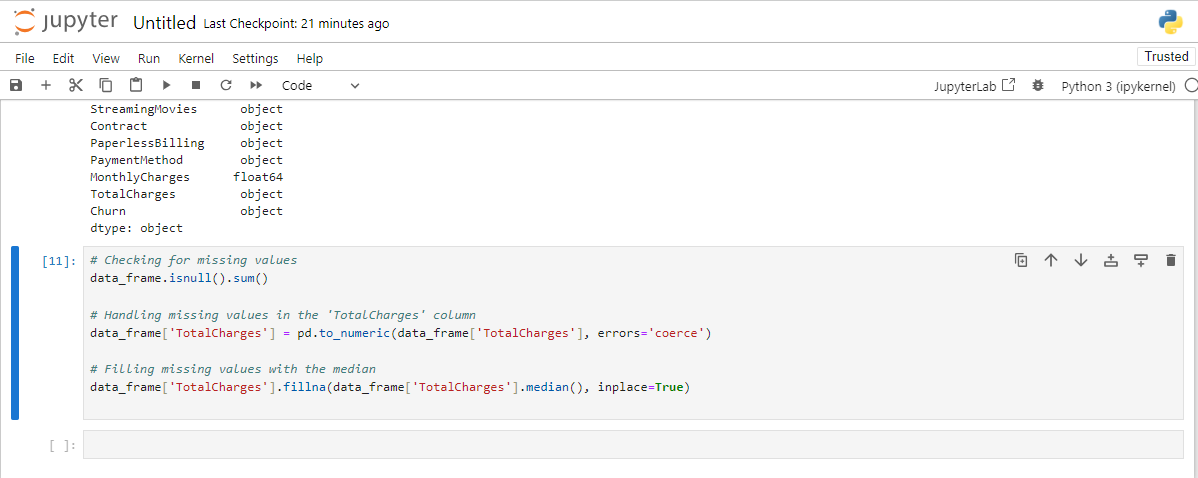


Figure 3: Jupyter Notebook Code and Output: Data Type Exploration and Handling Missing Values

* **Encoding Categorical Variables**: Gender, Contract and PaymentMethod would be categorical inputs that would have to be converted by the algorithm to numerical form for analysis.

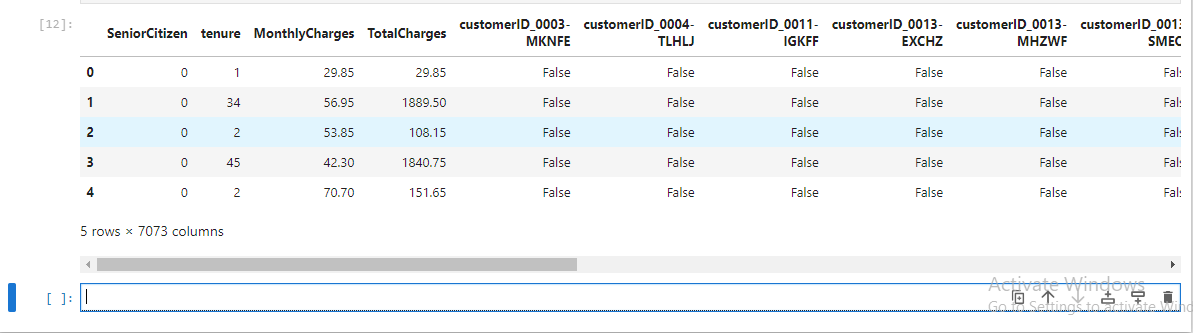


Figure 4: Jupyter Notebook Code and Output: Data Exploration and Handling Missing Values

## 2.4 Exploratory Data Analysis (EDA)

### 2.4.1 Descriptive Statistics

Also known as summarizing statistics, they let you understand the density and centrality of the data in terms of its features.

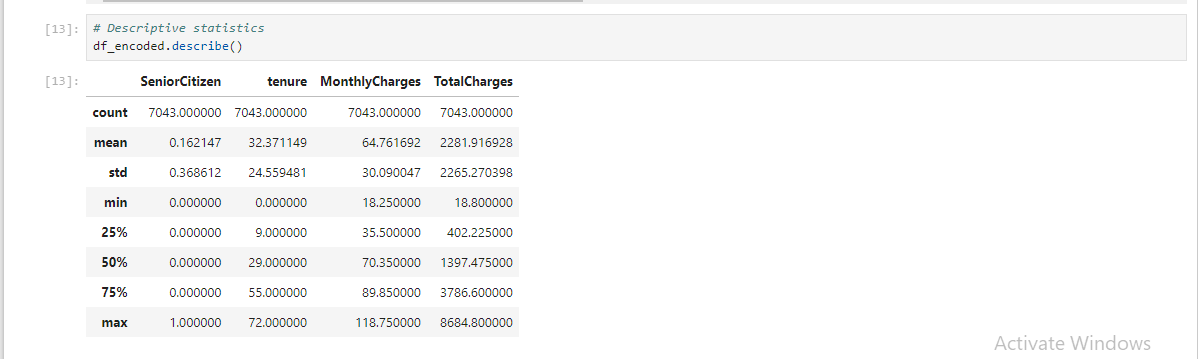


Figure 5: Jupyter Notebook Output: Descriptive Statistics for Numerical Columns

It is possible to derive certain qualitative conclusions from such statistical generalization; average tenure, MonthlyCharges, the percentage of SeniorCitizen customers, etc.

### 2.4.2 Data Visualizations

This means that, by presenting a view of the data, some insights can be made as to the interaction between different features and the target variable known as Churn.

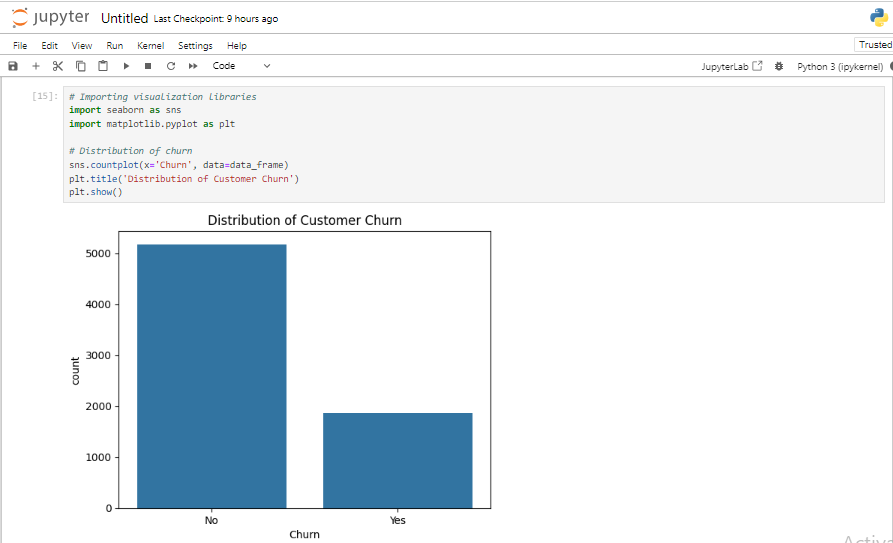


Figure 6: Jupyter Notebook Code and Output: Distribution of Customer Churn

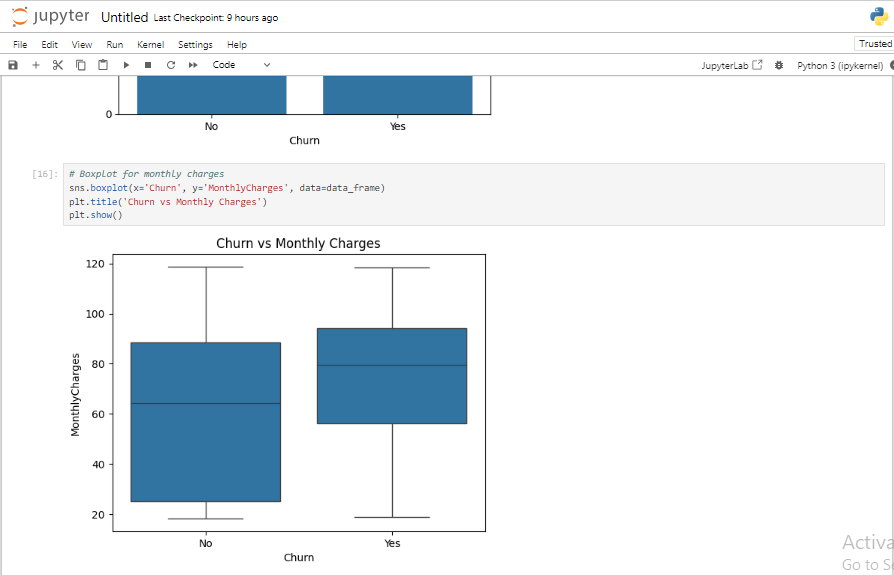


Figure 7: Jupyter Notebook Code and Output: Boxplot of Monthly Charges by Churn

* **Churn Distribution**: The count plot is a measure of the customers who have churned against those who have not.
* **Monthly Charges vs. Churn**: The boxplot gives a view as to how the monthly charges are spread out between the churning and the non- churning customers.
* **Correlation Heatmap**: This helps to visualize past correlations between different variables and shows how ‘strong’ each feature is with Churn.

### 2. 4. 3 Insights from EDA

1. **Churn Rate**: Currently, customer churn has been experienced within a significant percentage which underlines the need to come up with efficient methods of utilizing customer churn.
2. **Monthly Charges**: Mobile customers who are being charged a relatively high amount in a month are more likely to churn thus implying that improper pricing strategy might be a reason for customer churn.
3. **Service-Related Features**: Metrics like Contract, TechSupport and OnlineSecurity have a high degree of correlation with Churn implying that customer with shorter contractual period or who do not avail technical support are likely to churn. It would be expected that some of these variables to be one of the independent variables when developing common machine learning models.

These insights culminated from the EDA phase are very important especially when model building and evaluation is ongoing.

# 3. Task 2: Machine Learning Models

## 3. 1 Model Selection

Therefore, two ML models have been selected for performance of the task related to customer churn. They are ideal for classification problems, and they are prevalently applied in churn prediction.

### 3. 1. 1 Logistic Regression

Logistic Regression is linear model for binary classification that is applied for prediction of whether a customer is going to churn or not. It quantifies the propensity of a customer to churn by employing the input features and does a linear regression between the target and input features where the target variable is the log-odds of churning.

### 3. 1. 2 Random Forest Classifier

Random Forest is another example of ensemble learning method in which decision trees are generated in parallel and the results are combined in order to reduce the error. It is appropriate used in cases where there are a large number of data points with categorical and numerical variables and it can also model complex relationships in the data.

## 3. 2 Python Code Implementation

### 3. 2. 1 Data Preprocessing for Modeling

When the models will be applied from machine learning, the data set obtained must be divided into training set and testing set and the features must be normalized.



Figure 8: Jupyter Notebook Code and Output: Data Preprocessing and Model Training

## Model 1: Logistic Regression

The Logistic Regression model is executed with the help of LogisticRegression class which belongs to the sklearn library. The steps include fitting the model and checking the accuracy of the fitted model with the help of training data set and then testing it on the test data set.

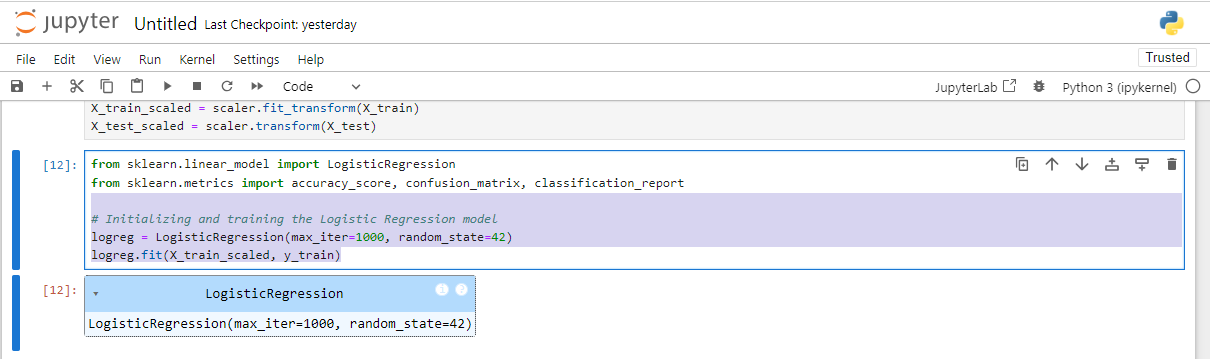


Figure 9: Jupyter Notebook Code: Logistic Regression Model Training

### 3.2. 3 Model 2: Random Forest Classifier

Random forest is also implemented using the RandomForestClassifier from sklearn. The training procedure of the model used in this case is similar to that of the Logistic Regression and the performance of both is compared.

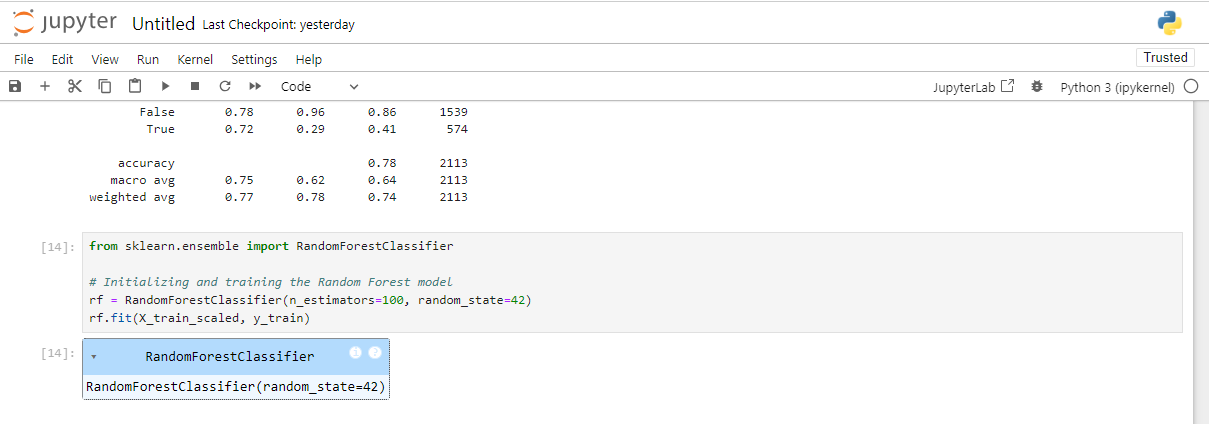


Figure 10: Jupyter Notebook: Random Forest Model Evaluation

### Explanation

* RandomForestClassifier(n\_estimators=100, random\_state=42): Initializes the Random Forest classifier with 100 estimators and fixes the random state for the same run multiple times.
* rf. fit(X\_train\_scaled, y\_train): Fit the Random Forest model on to the scaled training data.

Both models are trained and evaluated and in the next task, the different metrics shall be used to compare the two so as to determine which of the two models has the best results when it comes to customer churn prediction.

# Model Evaluation and Comparison

Here, two models of machine learning namely, Logistic Regression and Random Forest are being examined and compared based on various classification measures including, accuracy, precision, recall and F1-Score. These metrics give information on how the model are operating and is used to determine the best model which can be recommended for the business problem.

## 4. 1. Evaluation Metrics Used

* **Accuracy**: Calculates the accuracy of the model at large by defining the ratio of the dint of correct predicted values to the total values that the model was required to predict.
* **Precision**: Shows how many of the positive cases that have been predicted actually turned out to be positive.
* **Recall**: Makes the case of showing how many of the actual positive cases the model is able to capture.
* **F1-Score**: The average of precision and recall that is suitable where the classes are uneven, providing a balance of the two measurements.

## 4.2. Model 1: Logistic Regression – Performance and Visualization

After the training of the Logistic Regression model the model was tested using the test dataset. Below is the Python code used to evaluate and visualize the results:



Figure 11: Jupyter Notebook Code and Output: Logistic Regression Model Evaluation

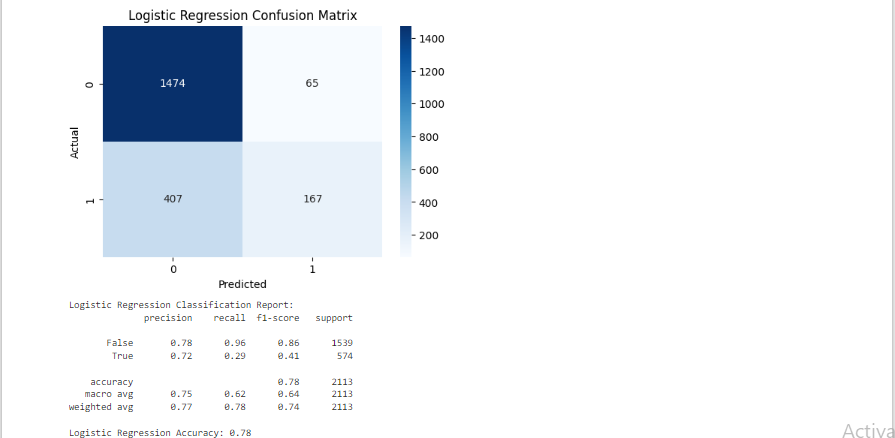


Figure 12: Logistic Regression Model Evaluation

Table 1: Logistic Regression Classification Report for Telco Customer Churn Prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| False | 0.78 | 0.96 | 0.86 | 1539 |
| True | 0.72 | 0.29 | 0.41 | 574 |
| Accuracy | - | - | 0.78 | 2113 |
| Macro Avg | 0.75 | 0.62 | 0.64 | 2113 |
| Weighted Avg | 0.77 | 0.78 | 0.74 | 2113 |

### Logistic Regression Performance Summary:

* Accuracy: 0. 78
* Confusion Matrix:
* True negatives (correctly predicted non-churn): 1474
* False negatives (incorrectly predicted churn): 407.
* True positives (correctly predicted churn): 167.
* False positives (incorrectly predicted non-churn): 65.
* Precision, Recall, and F1-score: The model produces a higher precision of 0. 78, and recall (0.96) for non-churn cases, but lower recalls (0.29) for churn cases, resulting in an overall F1-score of 0.41 for churn.

## 4.3. Model 2: Random Forest – Performance and Visualization

Another strong learner Random Forest was also used on the similar test dataset. Below is the code for evaluation and visualization:

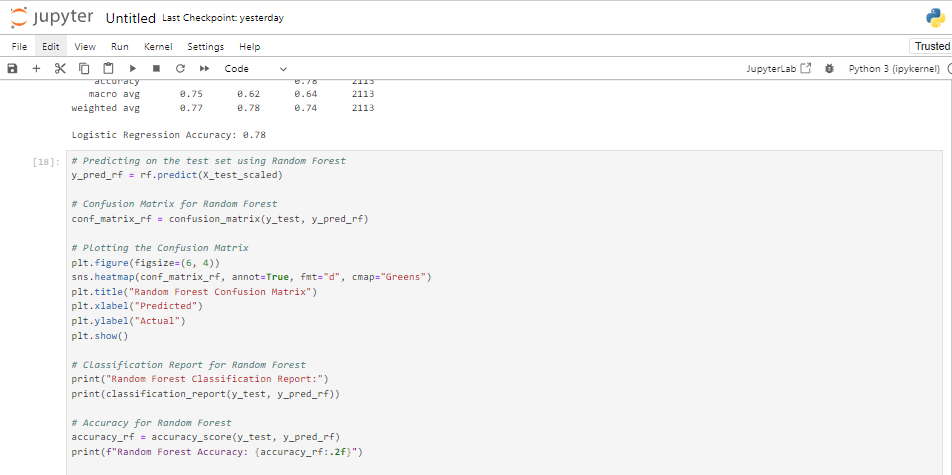


Figure 13: Jupyter Notebook Code and Output: Random Forest Model Evaluation

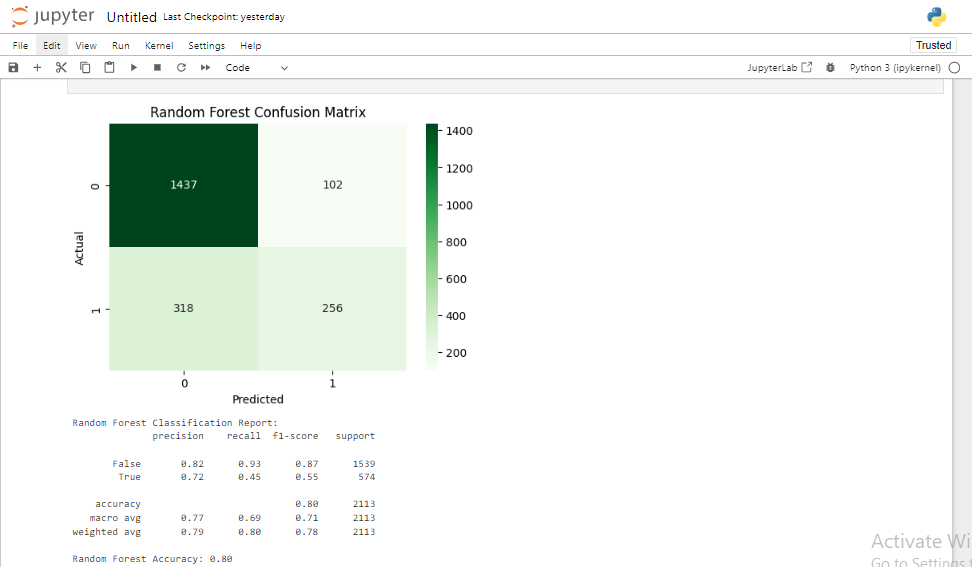


Figure 14: Jupyter Notebook Code and Output: Random Forest Model Evaluation

Table 2:Random Forest Classification Report for Telco Customer Churn Prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| False | 0.82 | 0.93 | 0.87 | 1579 |
| True | 0.72 | 0.45 | 0.55 | 574 |
| Accuracy | - | - | 0.80 | 2113 |
| Macro Avg | 0.77 | 0.69 | 0.71 | 2113 |
| Weighted Avg | 0.79 | 0.80 | 0.78 | 2113 |

### Random Forest Performance Summary:

* Accuracy: 0. 80
* Confusion Matrix:
* True negatives (correctly predicted non-churn): 1437
* False negatives (incorrectly predicted churn): 318
* True positives (correctly predicted churn):256
* False positives (incorrectly predicted non-churn): 102.
* Precision, Recall, and F1-score:
* Random Forest emerged with better recall, 0.45 for churn cases better than that of Logistic Regression. Again, the overall F1-score for churn cases was 0.55 which outperformed Logistic Regression Model moreover, it is better than the decision tree model with accuracy of 53 which has been discussed above.

## 4. 4. Analysis of Models and Identification of Model with Best Accuracy

There was also minor difference in the overall accuracy between the two models with Random Forest showing higher accuracy of 0.80 while Logistic Regression had 0.78 accuracy value. For the churn class, Random Forest had slightly higher recall values with 0.45 compared to 0.29 of Logistic Regression. The metrics of recall and F1-score are higher in Random Forest model and hence is a better model for solving this business problem of customer churn as it correctly identifies which customers are more likely to churn which is vital in customer retention.

## Recommendation

According to the identified performance criteria, Random Forest is deemed to be the most accurate model for churn prediction and provides a more nuanced strategy to identify instances of churn and non-churn. Since its recall is greater for churn cases, more of the potential churning clients are identified correctly, thus the retention strategies can be effectively implemented to maximum effect

# 5. Conclusion

## 5. 1. Summary of Key Findings

Using the Telco Customer Churn dataset it was possible to identify major tendencies that forecast customer behavior and the aspects that caused churn. When using a 10-fold cross validation it was noted that Random Forest yielded a slightly better accuracy of 80% compared to Logistic Regression’s 78% and a slightly better F1 score for the customer churn. Predictors related to the customers such as contract type, tenure, and the mode of payment were useful in determining churn behavior. Nonetheless, the recall outputs of both the models are fairly low for the positive churn class which gives an indication that there is a poor capability of the models in correctly identifying churners.

## 5. 2. Conclusion for Business/Social Problem

The implications of the study are the evidence for the fact that strengthening of customer retention efforts can be based on the particular factors including the type of the contract and payment methods. It can also be applied to customer segments with tenures, near the minimum of engagement, or using the month-to-month schemes which are the most vulnerable to churn. When it comes to the topic of structuring contact centers predictive models like Random Forest can determine the high-risk customers and companies can take actions before this happens.

## 5. 3. Future Directions and Potential Improvements

To improve the results of churn prediction, possible future works are presented below: Some more advanced algorithms can be used to increase the quality of the model, for example, Gradient Boosting or Neural Networks, which are likely to identify more detailed patterns of churn considering that earlier models could not accomplish this as effectively. Moreover, with regard to the imbalance in the data set, its treatment can be accomplished with such methods as SMOTE or, among others; the work on feature engineering can be effective. However, getting real time data feed and constantly updating the models could make churn prediction much more real-time and therefore make churn management more effective as models can be adapted mid-way, leading to much higher customer retention.

### 5. 3. 1. Model Improvements

* Applying hyperparameter tuning for getting the system into best performance.
* Applying other methods such as Gradient Boosting or Support Vector Machines instead of the current used ones for better predictive performance.
* Validation of the model particularly to prevent overfitting and under fitting of the model.

### 5. 3. 2. Additional Data and Features

* Having access to other fields of data that can be associated with the customer, e. g., history of interaction, or information from outside sources.
* Adding more features which are domain specific and which could potentially be indicative of command operations.
* Overcoming class imbalance in data to enhance the ‘Recall’ factor of the minority or the less represented classes using SMOTE.