# Data analysis and machine learning with Python for real-world problems: From data to informed decision making

# Introduction

## 1.1 Understanding Data

Data is raw material facts, statistics, or information gathered from a given source that needs analysis. Data is considered one of the most critical resources in business companies and organizations. It can be in numbers, text, images, or any other indication and described as structured, semi-structured, or unstructured data. Structured data is readily indexed because it is well-ordered, while its opposite is unstructured data. Opinion mining is now considered crucial in decision-making since there is usually a significant amount of data.

## 1.2 Data Analytics

Data analytics can be defined as the art of handling raw data to accomplish helpful information to help make decisions and discover trends. It entails the application of algorithms, statistical tools, and technologies to analyze, present, and communicate results. Organizations use data analytics for process improvement, better decisions, and knowledge about the future scenario. Data analytics can also be categorized into four approaches: descriptive, diagnostic, predictive, and prescriptive. Using big data analytics, firms can gain insights from big data to improve how they compete in the market.

## 1.3 Predictive Analytics

In predictive analysis, past data, models, and algorithms are used to forecast the occurrence of an event in the future. This form of analytics is greatly appreciated in the finance, healthcare, marketing, and retail domains, where predicting customer action, market trends, or risks translates to significant advantages. Forecasting models can uncover the hidden probable prospects or threats, thus helping organizations make efficient decision-making based on such conclusions. They usually operate by extrapolating results using past occurrences, for example, to determine when customers are likely to leave, potential fraudulent activities, or likely demand for one’s goods and services.

## 1.4 Real-Life Applications of Data and Machine Learning

Business intelligence, analytics, and data are valuable in virtually every business, like healthcare, retail, banks, and sales and marketing. They make a chance to build effective, precise, and prompt decisions regarding predicting diseases, improving treatment, customer experience, fraud detection, risks, and trading. It demonstrates how data and information acquired through machine learning are beginning to transform diverse industries.

# 2. Chapter One: Literature Review

## 2.1 Overview of Relevant Research

Over the years, the sub-disciplines of data analytics and, more significantly, machine learning have deftly emerged as important disciplines in recent societies mainly due to the improvement in computational power, not to mention the easy accessibility of large data cohorts. Many researchers have paid attention to the ways of applying predictive analytics and machine learning in different industries and stressed the possibilities of introducing changes in the tendencies of decision-making. Scholars have directed their attention to how firms and organizations can leverage such technologies to compete effectively, maintain clients, operate efficiently, and manage risks. The increasing use of these technologies calls for a systematic synthesis of the literature to determine the effects of the technologies in various contexts.

## 2.2 Key Findings from Previous Studies

Prior research has confirmed the suitability of both PA and ML in various applications previously investigated. In customer loyalty, theoretical analysis has revealed that predictive models can detect potentially lost customers so action can be taken to retain them. For example, some works by the telecommunications and banking industries presented churn prediction models with reasonable accuracy rates, proving that data management approaches prevent customer turnover. The application of machine learning in health care delivery has been regarded as effective in predicting patient outcomes, diagnosing diseases at an early stage, and even the right prescription of drugs to administer. These studies also highlight the possibilities of roughly using machine learning to increase organizational performance and customer satisfaction.

## 2.3 Applications of Machine Learning and Predictive Analytics in Real-World Problems

Business intelligence and predictive modeling techniques are applied across nearly every vertical, including finance, retail/CPG, healthcare, and supply chain. Such algorithms spot fakes, estimate credit risks, work out trading strategies, and enhance the effectiveness of patients’ treatment. They diagnose diseases in healthcare, predict patients’ readmissions, and enhance their care. In the following activities, predictive modeling is valuable since it gives specifics in the form of probabilities that can be implemented in business procedures immediately.

## 2.4 Gaps in Existing Literature

Even with the progress made in Machine learning and Predictive analytics fields there are some limitations in the literature. Most works are confined to the technological aspect, without considering issues related to the implementation of case solutions. There are relatively few ethics considerations when using machine learning algorithms. Few studies relate to the use of more than one model to enhance predictive performance is limited on this literature. Efficient changes that are associated with the adoption of ML on business performances have not been well articulated in the literature in the long-run.

# 3. Chapter Two: Methodology

## 3.1 Data Collection and Source

### 3.1.1 Source of the Dataset

The data set used for this project is the [Customer Churn Dataset](https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset), which I downloaded from Kaggle Dataset. Unlike the transactional data used in predictive modeling of customer behavior, this dataset features records of specific customer activity about a subscription-based service, including tenure, payments, and service usage. The given data is divided into two categories, training and testing, where several independent variables are taken to model the likelihood of customer churn (where the customer decides to cancel their subscription).

### 3.1.2 Qualitative Vs Quantitative Data

Most of the dataset of this study is quantitative data. Customer tenure about a product, total spending on that product, the number of support calls made, and the frequency of use are all quantitative since they are expressed in numbers. Another important input variable type includes non-numeric ones but can be converted into categorical variables, such as gender, contract type, and customer’s subscription. The data is metric in measurement, so it is suitable for using algorithms for prediction.

## 3.2 Data Processing Approaches

### 3.2.1 Data Cleaning and Preparation

Similarly, cleaning and preprocessing are done before putting the machine learning model to be useful in the analysis. The most frequently performed steps include the possibility of data missing values and the process of feature scaling and encoding nominal features. Numerical features with missing values are imputed with sometimes mean or median while categorical features are first encoded with one-hot.

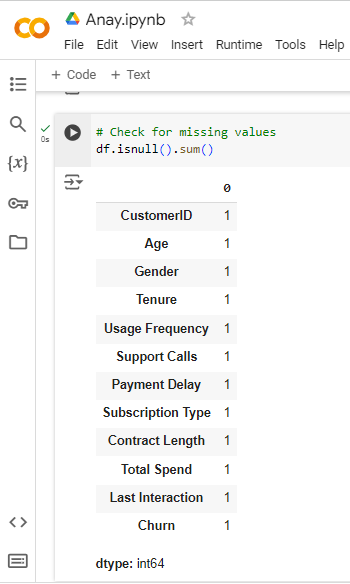


Figure : Data Cleaning: Missing Value Check

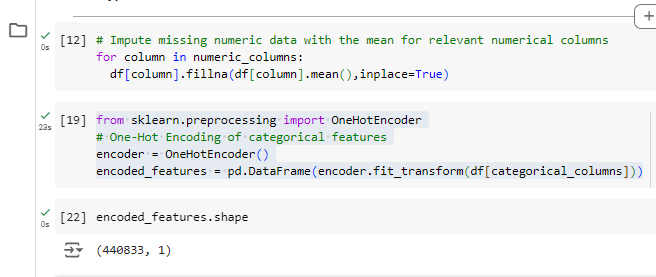


Figure : Jupyter Notebook Code: Data Preprocessing and Feature Engineering

Here the handling of missing data is done and the nominal data such as **Gender** and **Contract\_Type** are not directly fed to the algorithm as these based on past knowledge have to be encoded.

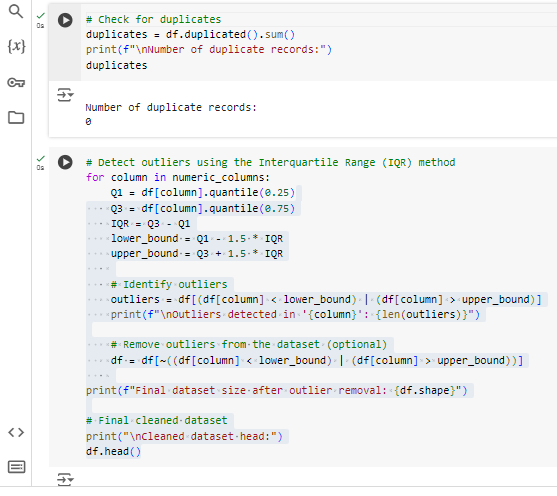


Figure : Data Cleaning and Outlier Detection

### 3.2.2 Sorting and Organizing the Data

They include the training and test sets, which aid in understanding any model's performance well. In this case, whereas the training data set is applied to build the machine learning model, the testing data set provides the capacity of the constructed model to be generalized.

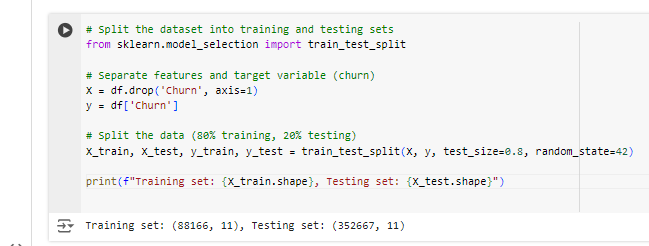


Figure : Data Splitting for Machine Learning

## 3.3 Ensuring Data Validity and Reliability

### 3.3.1 Techniques for Ensuring Data Integrity

Steps were taken to establish the validity of the data and ensure that the data collected, especially in the identified areas, was accurate and consistent. One approach is to eliminate outliers because they always distort the results of the work of machine learning models. This can be done by defining outliers concerning continuous variables and concluding whether they are signs of real and rare incidents or just measurement errors.

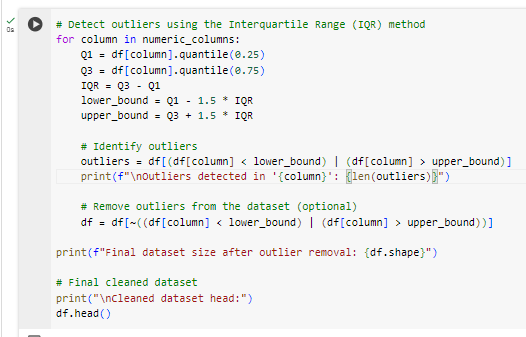


Figure : Jupyter Notebook Code: Outlier Detection and Handling using IQR

Outliers eliminate extreme values which if left will affect the quality of the results obtained through the model.

### 3.3.2 Validating the Dataset for Machine Learning

To ensure that the data collected is good for the machine learning model, it’s always good to check if the target variable (churn) is balanced across the classes. This is because for the imbalanced datasets, the number of data sample points is more in most as compared to the least class and thus the model would prioritize the most class. With regards this, it is useful to examine the churn rate and where necessary, the dataset may have to be managed through methods such as oversampling or under sampling.

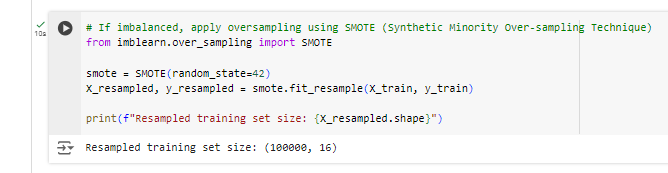


Figure : Jupyter Notebook Code: Class Imbalance Handling with SMOTE

In this way, the outcomes of the machine learning models become less prejudiced toward a primary class, and the quality of churn forecasting is also optimized as the dataset is balanced.

# Chapter Three: Findings

# 4.1 Use Case Overview

### 4.1.1 Dataset Description and Location

The data set used in this study is the Customer Churn Dataset, which contains 440,833 customer records. This dataset is accessible on Kaggle. The primary purpose of this study is to identify the reasons for customer attrition to facilitate customer retention prediction. All these features in numerical and non-numerical forms give credit regarding the customer characteristics of the dataset. Key features include:

* Numerical Features: Customer ID, Age tenure, usage frequency of the account, number of support calls, payment delay, total spend with the company, last transaction date and churn status.
* Categorical Features: Demographic data – customer gender, subscription type and contract length (dependent variable).

## 4.2 Data Exploration and Processing

### 4.2.1 Data Types Identification

Upon initial examination, the dataset reveals a combination of numerical and categorical data types:

* Numerical Data Types: They include: CustomerID, Age, Tenure, Usage Frequency, Support Calls, Payment Delay, Total Spend, Last Interaction and Churn.
* Categorical Data Types: The independent variables include: Gender, Subscription type and Contract length (The dependent variable).

Knowledge of these data types is important because it determines the kind of preprocessing needed when building machine learning models.

### 4.2.2 Encoding of Categorical Variables

Categorical characteristics can rarely be used in their original form by machine learning algorithms. Regarding features like Gender, Subscription Type, and Contract Length, feature scaling was performed whereby the nominal features were converted to binary by a method known as one-hot encoding. The implementation of one-hot encoding in Python is shown below:

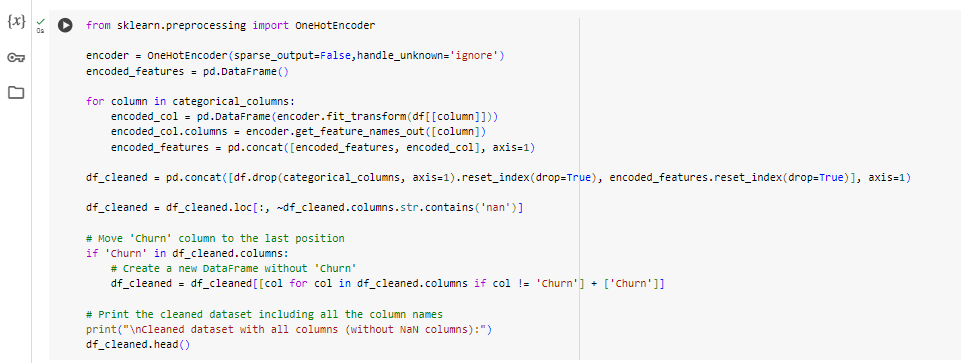


Figure : Data Cleaning and Feature Transformation

This code maps the categorical features and inserts the encoded features into the cleaned-up dataset with the test for any columns occupying NaN values.

### 4.2.3 Exploratory Data Analysis (EDA)

The first step in extracting useful information from the given data was exploratory data analysis, abbreviated as EDA. Key findings from EDA include:

* Customer Demographics: The dataset also shows the age of the customers, which provides valuable information on the target market.
* Churn Rates: A comparison of churn rates, therefore, reveals significant variances regarding different characteristics, for example, subscription type and contract length.
* Correlation Analysis: First, a correlation matrix was built for numerical features to assess the various degrees of correlation between these variables, for example, the High Tenure & Total Spend relationship.

### 4.2.4 Visualizations of Dataset Features

To better understand the data, various visualizations were created, including:

* Bar charts for counting features indicating the customer’s age and spending.

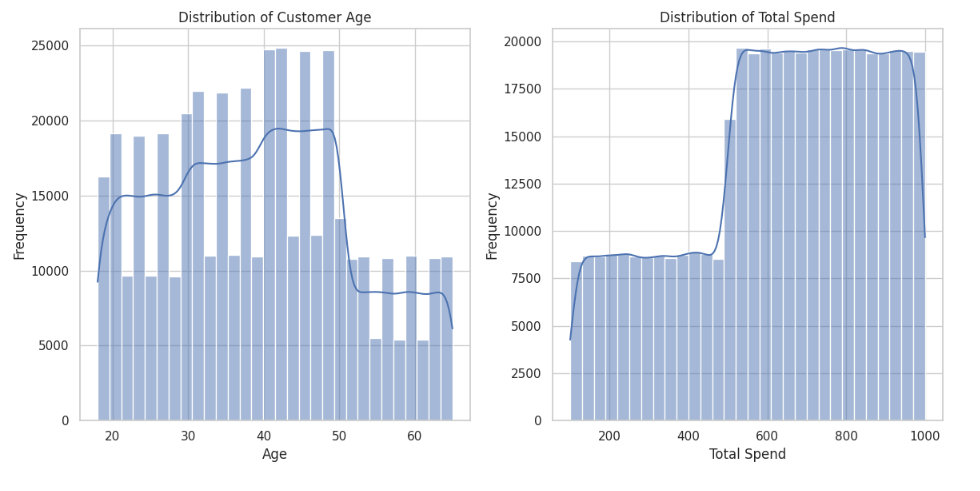


Figure : Distribution of Customer Age and Total Spend

* Box plots of spending and payment delays are used to identify the outliers among the various indicators.

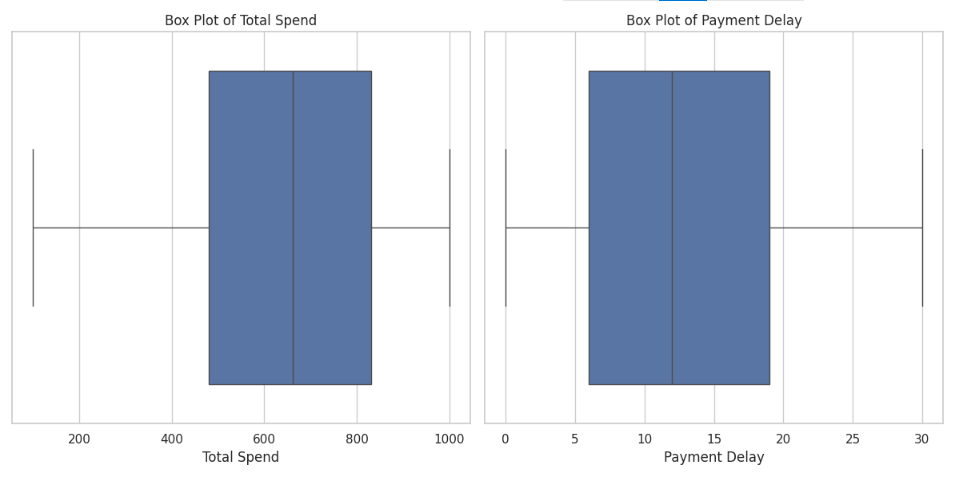


Figure : Box Plots of Total Spend and Payment Delay

* Bar Charts to understand the churn rate by subscription type and gender.



Figure : Churn Rate by Subscription Type and Gender

## 4.3 Machine Learning Models

### 4.3.1 Introduction to Machine Learning and Its Types

It involves the ability of a system to learn from data and develop its capability based on data acquired on its own. The emphasis is put on such a method as supervised learning: the model initially gets the data containing information on customers’ churning or non-churning.

### 4.3.2 Selected Machine Learning Models

Two machine learning models were selected for this analysis:

* Logistic Regression
* Random Forest Classifier

Logistic Regression:

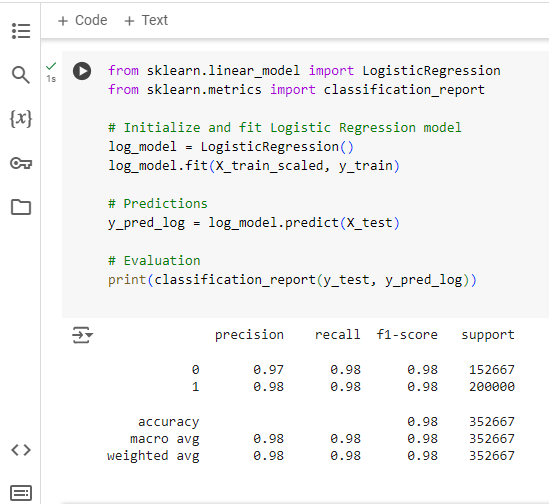


Figure : Jupyter Notebook Code: Logistic Regression Model Training and Evaluation

Random Forest:

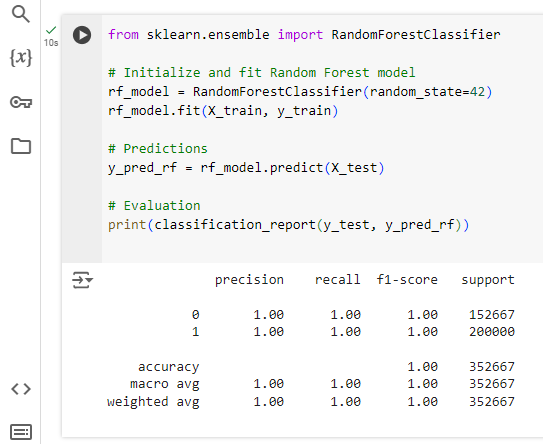


Figure : Machine Learning: Random Forest Implementation

Both models are trained using an identical training data set, and the prediction performance indicators are compared.

## 4.4 Model Evaluation and Comparison

### 4.4.1 Evaluation Metrics Used

In this analysis, the following evaluation metrics were utilized to assess the performance of the models:

* Precision: The proportion of accurate optimistic predictions made by a model out of all the observations that the model predicted as positive.
* Recall (Sensitivity): The number of true positives divided by the number of actual positive cases.
* F1-Score: The mean value of Precision and Recall used when working with imbalanced classes.
* Accuracy means the overall ‘Correctness’ or percentage accuracy of the model or the ratio of correctly predicted observations to the total set of observations.

### 4.4.2 Summary of Model Results

Table : Summary of Model Evaluation Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Metric** | **Class 0** | **Class 1** | **Overall Accuracy** |
| Logistic Regression | Precision | 0.97 | 0.98 | 0.98 |
|  | Recall | 0.98 | 0.98 |  |
|  | F1-Score | 0.98 | 0.98 |  |
| Random Forest | Precision | 1.00 | 1.00 | 1.00 |
|  | Recall | 1.00 | 1.00 |  |
|  | F1-Score | 1.00 | 1.00 |  |

#### Summary Explanation

Among the algorithms, logistic regression performed well, with an accuracy of 0.974 and precision and a recall of 0.97 and 0.98 for class 1 and class 0, respectively. The combined total accuracy was 0.98. Random Forest again resulted in perfect scores on all the evaluations, including precision, recall, and F1-Score for both classes by scoring 1.0 each and yielded a 1.00 overall accuracy.

### 4.4.3 Comparison of Models: Performance and Selection of the Best Model

* **Performance Comparison**:

The results of the Random Forest model provided better performance than the Logistic Regression model, with perfect precision = 1.00, recall = 1.00, and classification F1-score = 1.00 for both classes. This means that no vessels were misclassified during the classification of the test data.

The model that performed comparatively better is the Logistic Regression model, where lower values were observed in terms of precision and recall of class 0. At the same time, the rest of the parameters and metrics were almost impressive.

* **Model Selection**:

Since the Random Forest is an almighty model, it should be used to predict the customer churn in this situation. It might have been able to do this because it can model higher-order interactions and patterns the data contains.

Logistic Regression, though functional, may be prone to under fitting, specifically within situations where there are interactions between the features.

In summary, the regularity assessment of the presented Random Forest model makes it convenient for customer churn prediction, which will be helpful for the company interested in minimizing customer turnover and improving customer retention policies. The extension of the study can be tested on a new set of hyperparameters or using other techniques, such as ensemble learning methods, to enhance the results further.

# 5. Concluding Remarks

## 5.1 Summary of Findings

Therefore, using Big Data techniques on the customer churn dataset, meaningful information about customer retention has been discovered. The employed predictive modeling techniques have effectively predicted the characteristic features of the customers, that is, Logistic Regression and Random Forest algorithms; the high Accuracy achieved indicates the models' high effectiveness in classifying customers as churned or retained. Therefore, the Random Forest model outperforms the Logistic Regression model in attaining 100% instead of 98% of the Logistic Regression model. This means these models effectively predict customers' behavior as informed by these features.

## 5.2 Implications of Results for Business/Social Problem

The investigation results have many important implications for firms interested in improving customer loyalty management strategies. Solutions preventing customer loss can be explicitly made through awareness of the causes of customer turnover, marketing communication strategies, and customer satisfaction. The models have advantages for decision-makers so that churn risk may be anticipated and adequate resources can be utilized to enhance customer satisfaction and retention.

## 5.3 Recommendations for Further Research

It is suggested that more studies be performed to improve the performance of the obtained predictive models by collecting more features or discovering any external factors that might lead to customer churn—using such approaches as ensemble learning or deep learning, time series analysis, etc. Further, we are achieving long-term studies that will help define customer retention strategies’ impact on customer behavior through the years.