جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ ⵏ ⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

كلية العلوم والتقنيات فاس
ⵜⴰⵖⵉⵡⴰⵏⵜ ⵏ ⵜⵎⴰⵙⵙⴰⵏⵉⵏ ⴷ ⵜⵙⵏⵉⵅⵉⵜⵉⵏ - ⴼⴰⵙ
FACULTÉ DES SCIENCES ET TECHNIQUES DE FÈS

Projet de fin d'étude

Licence sciences et techniques Génie Informatique

Exploratory Data Analysis of a telecommunication company's customers to predict churn

Laboratoire Systèmes Intelligents & Applications
LSIA

Lieu de stage: Laboratoire Systèmes Intelligents et Applications FST FES.

Réalisé par :	Encadré par :
Soukaina RHAZZAFE	Pr. Aicha MAJDA

Soutenu le 09/07/2021 devant le jury composé de :

Pr. Aicha MAJDA

Pr. Mohamed OUZARF

Pr. Abdelali BOUSHABA

Année universitaire 2020-2021

Acknowledgments

I would like to thank God first, as with his blessings and grace that I have been able to bring this project to light.

I cannot express enough thanks to my supervisor **Professor Aicha MAJDA** for fueling my enthusiasm by giving me the opportunity and the chance to carry out such an interesting project, for helping me throughout it by the valuable guidance and advice.
I will be forever grateful for you.

I would like to thank the jury members **Professor Mohamed OUZARF** and **Professor Abdelali BOUSHABA** for accepting to evaluate this project.

I would like to thank also the academic personnel, especially the department of IT, and the administrative staff of the University for their enormous efforts to provide a quality education.

Special thanks to my parents and brother for the constant motivation and unconditional support and encouragements through all these years.

Abstract

With the exponential growth in generated data and the great value it holds, analyzing it to extract meaningful insights and base important decisions on them has become a necessity, especially for companies as it is now crucial for their success.

Customers churn is one of the most common and costly issues with telecom companies, one way to address this issue is to predict new customers' decision by analyzing old one's data in order to understand why they chose to churn and on what metrics they based this decision and then react upon the results.

The aim of this end-of-study project was to use data analytics tools to understand and analyze data provided by a telecom company with a moderately high churn rate and also to put in use Machine Learning models and choose suitable ones to deploy in a Web application that predicts customers' decisions.

Résumé

Avec la croissance exponentielle de la création de données et la valeur qu'elles rapportent leur analyse dans le but d'extraire des connaissances significatives sur lesquelles se basent des décisions importantes est devenue une nécessité, notamment pour le succès des entreprises.

Le désabonnement des clients est l'un des problèmes les plus courants et couteux pour les sociétés de télécommunication, une façon de traiter ce problème est de prévoir le désabonnement des nouveaux clients en analysant les données des anciens pour comprendre pourquoi ils choisissaient le désabonnement et la base de leur décision pour que la société puisse réagir.

Ce projet de fin d'études avait comme but l'utilisation des outils d'analyse des données d'une société, qui a un taux de désabonnement un peu élevé, et des modèles d'apprentissage automatique pour choisir ceux qui sont convenables au déploiement sur une application Web qui prévoit le désabonnement des clients.

Table of contents

List of figures

List of abbreviations

AI	**A**rtificial **I**ntelligence
ANN	Artificial Neural Network
API	**A**pplication **P**rogramming **I**nterface
CRISP	**C**ross**-I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
CSS	**C**ascading **S**tyle **S**heets
CSV	**C**omma **S**eparated **V**alues
EDA	**E**xploratory **D**ata **A**nalysis
HTML	**H**yper**T**ext **M**arkup **L**anguage
IDE	**I**ntegrated **D**evelopment **E**nvironment
MLP	**M**ulti **L**ayer **P**erceptron
NB	**N**aïve **B**ayes

Introduction

This project was based on a two months long end-of-study internship, included in the Faculty of Sciences and Technology FES's "License sciences et technique Génie Informatique" program that I have been given a golden opportunity to carry on at the LSIA Laboratory, a research laboratory in the same university. Its aim was answering a real life problematic, a telecommunication company's moderately high customers' churn rate, by utilizing data analysis tools and techniques as well as Machine Learning models to predict customers' decision and then deploying the suitable ones in a Web application.

This report provides an overall presentation of the project and its workflow it includes three chapters:

The first chapter, *General context of the project*, provides a presentation of the LSIA laboratory, where the internship has taken place, an overview of the problematic and the objective of this project as well as the proposed solution.

The second chapter*, Predictive Modeling: EDA and ML*, introduces the concepts of Data Analytics in general and Exploratory Data Analysis in particular, their tools and techniques, and Machine Learning.

The third chapter, *Customer Churn Prediction Application*, includes two section where the first is a presentation of tools used in the analysis and modeling phase, and the second presents those used in the deployment of the models and development of the churn prediction application.

Chapter 1: General context of the project

I. Host organization

LSIA laboratory (Laboratoire des Systèmes Intelligents & Applications) was first accredited in 2011 as a Science and Engineering Techniques Doctoral Studies Centre (CED) at the Faculty of Science and Technology Fes by the Sidi Mohammed Ben Abdellah University.

Composed by 20 professors and 40 doctoral students, LSIA Laboratory's main ambition is to make scientific and technical progress, and then base upon it solutions of concrete problems, in the different fields of IT.

The SIS project (Systèmes Intelligents au service de la Société), the main project of the laboratory for the accreditation period 2020-2025, is aimed at promoting, facilitating and coordinating all research activities of the members in order to set up Smart Solutions in the fields of: E-Health, Intelligent Transportation Systems…

Fields of research:

- Artificial Intelligence, Machine Learning , Multiple-criteria decision analysis
- Data Mining and Decision Making, Data Warehouse, Big Data
- Image Processing, Speech Processing, Language Processing
- Embedded systems , Code theory
- Ad hoc networks, networks security
- E-Learning, E –health

Collaborations:

- Picardie Jules Verne University (UPJV) Amiens France
- Montpelier University
- UHC FES (CHU)
- Limerick University, Ireland
- Berger-Levrault Montpelier Company, France

II. Study of the existing

As the role of technology becomes vital in each and every sector, global data sources have rapidly increased creating huge amounts of information every day, as over the last two years alone 90 percent of the data in the world was generated.

With about 1.7 megabytes worth of data created every second for every human being on the planet, a whopping total of 64 zettabytes- a zettabyte equals 1000^7 (1,000,000,000,000,000,000,000) bytes- has been generated in 2020 surpassing the estimated 40 zettabytes. This astonishing growth has no sign of slowing down, as it is predicted to reach 180 zettabytes by 2020. [1]

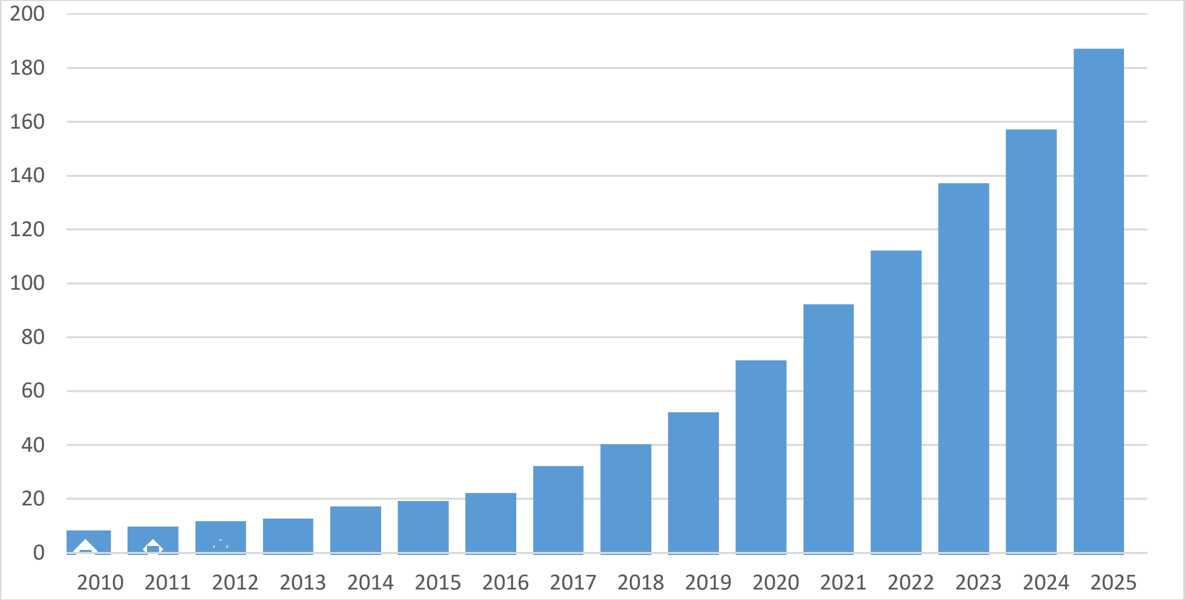


Figure 1: The exponential data growth between 2010 and 2020 in Zettabytes

The bulk of this data comes from three primary sources: social data, machine data and transactional data. [2]

- Social data: Likes, Tweets & Retweets, Comments, Video Uploads, and general media that are uploaded and shared via social media platforms.
- Machine data: generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behavior.
- Transactional data: generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts…

In 2017, the "Economist" claimed that data replaced oil as the world's most valuable resource, **"A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These titans—Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft—look unstoppable. They are the five most valuable listed firms in the world. Their profits are surging: they collectively racked up over $25bn in net profit in the first quarter of 2017."** [3]

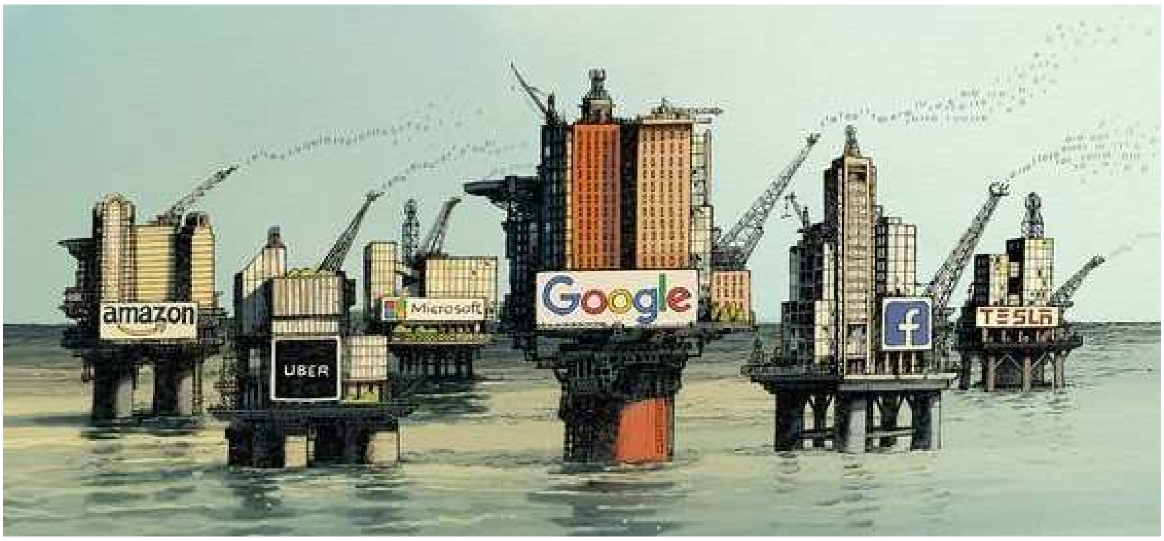


Figure 2: Data is the new oil

But what would all this data hold that makes it such a valuable resource? Well, hidden in it are business insights that can trigger explosive growth - social data for example provides invaluable insights into consumer behavior…. insights that can create business value and be enormously influential in many fields, and even beyond insights are actionable insights, those that drive action. [4]

Extracting actionable insights from data has become a necessity for companies in this "data-driven" age, to make efficient decisions, it helps them determine their positions in the market relative to their competitors and identifying the potential risks that need to be avoided and the opportunities that must be taken in order to grow, therefore it has become crucial to companies' success.

This has led to a growth in the global Data Analytics- the science of analyzing and processing raw data to extract meaningful insights- market that has been estimated to reach 24.63 billion USD in 2021, especially as the COVID-19 pandemic that has accelerated the adoption of data analytics solutions and services globally. [5]

Despite the huge amount of generated data and the growing need to analyze it, less than 1% of it is analyzed. [1] By further researching data analytics statisticians have discovered that not all generated data has the potential to bring value or hold meaningful insights. Also data is not all created equal; data generated from social media applications is completely different from the data generated by machines, it can be sorted into one of two categories: structured and unstructured data. Structured data is data that has been predefined and formatted to a set structure before being placed in data storage, while unstructured data is data stored in its native format and not processed until it is used. Most of the data we generate today is unstructured, which means it comes in different forms, sizes, and even shapes, which makes it difficult and costly for companies to manage and analyze it. And needless to say, sifting through all of that data, parsing it - converting it into a format more easily understood by a computer, and analyzing it is way too much for human minds to tackle, artificial intelligence's algorithms have been written to accomplish the task of deriving insight out of complex data with all its forms and structures. [6]

Artificial Intelligence tools, equipped with machine learning, are introducing numerous ways to process, analyze data and extract insights that help improve Decision Making by providing predictions based on the given data.

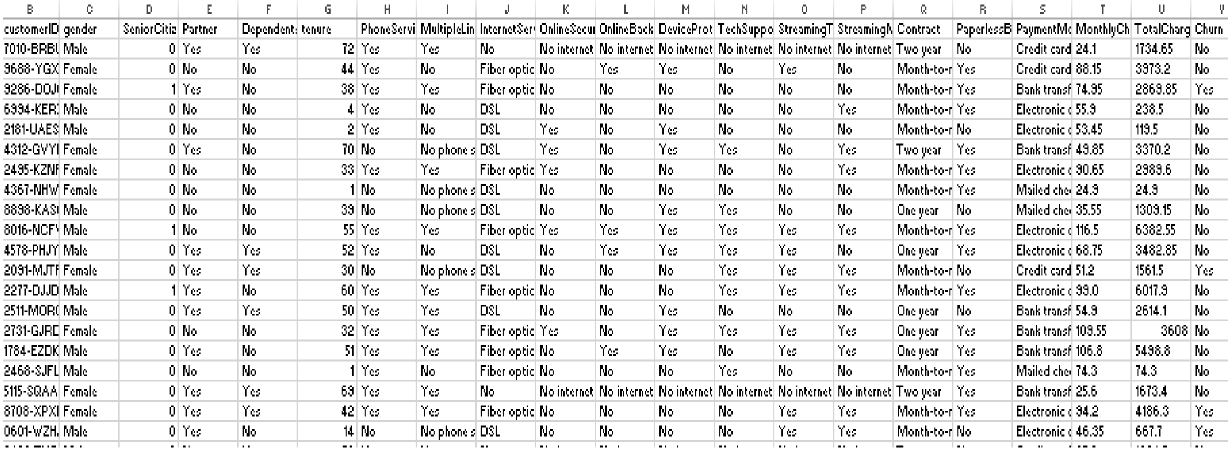
It was in this context that I was given this end-of-study project at the LSIA laboratory, where the main goal was to not only have a solid grasp of the concept of data analytics and its tools, processing and analyzing data and their techniques, but also learn how to put in good use the Artificial Intelligence's algorithms for effective Decision Making.

The data on which that I will be performing processing and analysis techniques was provided by an American telecommunication company by Kaggle, the biggest data science community in the world. It is a platform that allows users to find free real-time datasets or publish them, usually in a csv format, explore and build Machine Learning models in a web-based environment, work with other data scientist, and enter competitions to solve data science challenges.

The dataset provided is in csv format where each row represents a customer and each column contains one of the company's customer's attributes kept in records.

It includes information about:

- Demographic information about customers: gender, age range, and if they have partners and dependents.
- Customer account information: tenure- how long they've been a customer in months-, contract, payment method, paperless billing, monthly charges, and total charges.
- Services that each customer has or hasn't signed up for: phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customers who left within the last month: the column is called Churn.



B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
customerID	gender	SeniorCitiz	Partner	Dependent	tenure	PhoneServi	MultipleLin	InternetSer	OnlineSecur	OnlineBack	DeviceProt	TechSuppo	StreamingT	StreamingM	Contract	PaperlessB	PaymentMe	MonthlyCh	TotalCharg	Churn
7010-BRBU	Male	0	Yes	Yes	72	Yes	Yes	No	No internet	No internet	No internet	No internet	No internet	No internet	Two year	No	Credit card	24.1	1734.65	No
9688-YGX	Female	0	No	No	44	Yes	No	Fiber optic	No	Yes	Yes	No	Yes	No	Month-to-r	Yes	Credit card	88.15	3973.2	No
9286-DOJ	Female	1	Yes	No	38	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to-r	Yes	Bank transf	74.95	2869.85	Yes
6994-KER	Male	0	No	No	4	Yes	No	DSL	No	No	No	No	No	Yes	Month-to-r	Yes	Electronic c	55.9	238.5	No
2181-UAES	Male	0	No	No	2	Yes	No	DSL	Yes	No	Yes	No	No	No	Month-to-r	No	Electronic c	53.45	119.5	No
4312-GVYI	Female	0	Yes	No	70	No	No phone s	DSL	Yes	No	Yes	Yes	No	Yes	Two year	Yes	Bank transf	49.85	3370.2	No
2495-KZNF	Female	0	No	No	33	Yes	Yes	Fiber optic	Yes	No	No	No	No	Yes	Month-to-r	Yes	Electronic c	90.65	2989.6	No
4367-NHW	Female	0	No	No	1	No	No phone s	DSL	No	No	No	No	No	No	Month-to-r	Yes	Mailed che	24.9	24.9	No
8898-KASI	Male	0	No	No	39	No	No phone s	DSL	No	No	Yes	Yes	No	No	One year	No	Mailed che	35.55	1309.15	No
8016-NCFY	Male	1	No	No	55	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Month-to-r	Yes	Electronic c	116.5	6382.55	No
4578-PHJY	Male	0	Yes	Yes	52	Yes	No	DSL	No	Yes	Yes	Yes	Yes	No	One year	Yes	Electronic c	68.75	3482.85	No
2091-MJTF	Female	0	Yes	Yes	30	No	No phone s	DSL	No	No	No	Yes	Yes	Yes	Month-to-r	No	Credit card	51.2	1561.5	Yes
2277-DJJD	Male	1	Yes	No	60	Yes	Yes	Fiber optic	No	No	No	Yes	Yes	Yes	Month-to-r	Yes	Electronic c	99.0	6017.9	No
2511-MORC	Male	0	Yes	Yes	50	Yes	Yes	DSL	No	No	Yes	No	No	No	One year	No	Bank transf	54.9	2614.1	No
2731-GJRD	Female	0	No	No	32	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	One year	Yes	Bank transf	109.55	3608	No
1784-EZDK	Male	0	Yes	No	51	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	One year	Yes	Bank transf	106.8	5498.8	No
2468-SJFL	Male	0	No	No	1	Yes	No	Fiber optic	No	No	No	Yes	No	No	Month-to-r	Yes	Mailed che	74.3	74.3	No
5115-SQAA	Female	0	Yes	Yes	69	Yes	Yes	No	No internet	No internet	No internet	No internet	No internet	No internet	Two year	Yes	Bank transf	25.6	1673.4	No
8708-XPXI	Female	0	Yes	Yes	42	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to-r	Yes	Electronic c	94.2	4186.3	Yes
0601-WZH	Male	0	Yes	No	14	No	No phone s	DSL	No	No	No	No	Yes	Yes	Month-to-r	No	Electronic c	46.35	667.7	Yes

Figure 3 : The dataset provided by the company in csv format

The given task is to analyze this data and predict the churn of the company's customers, which makes it a classification problem, since the customers will be classified by their decision to leave or not.

Reading the dataset from the excel file, as figures, is not easy, it contains so many rows and columns and this format makes it hard to understand the data, identify patterns and find correlations between the attributes and remove redundancies to learn more about the company's customers who decided to leave it.

III. The proposed solution

Customer churn costs the company a huge amount in lost revenue every year, as acquiring a new customer costs five times as much as keeping an existing one [7], so a reduction in customer churn would be a smart move for it to avoid huge losses and maybe even increase their revenues in an ever-competitive business environment, that's where churn prediction comes into play, it allows the company to understand on what metrics the customers decide to leave, and then react upon them.

Machine Learning will be the best option to automate the task of customers churn prediction, as AI based tools have the analytical capability to process large amounts of data within minutes to provide meaningful business-based insights. And because it is a classification task, ML classification models and algorithms should be used.

But before being able to utilize these algorithms, it is important to understand how to initially get the data in proper order and how to effectively analyze the data in order to produce the best model. Exploratory Data Analysis is primarily for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

In this project, I will be making an application that reads the dataset in the csv format and performs the necessary processing and analysis tools to better visualize the data in the aim of getting insights and patterns regarding why customers choose to leave or not and then be able to predict, with a "good" accuracy, the customers' decisions.

In this report, I will be introducing different techniques of Exploratory Data Analysis and visualization and then classification models for decision making, along with the used tools such as the python programming language and its AI libraries.

The next chapter contains in deep details of these techniques and tools used to develop and set up the application.

To carry out this project, I followed a set of 5 general steps of a predictive modeling project, based on a CRISP - Cross-Industry Standard Process, these steps can be followed sequentially as much as going back and forth between them as needed is possible. [8]

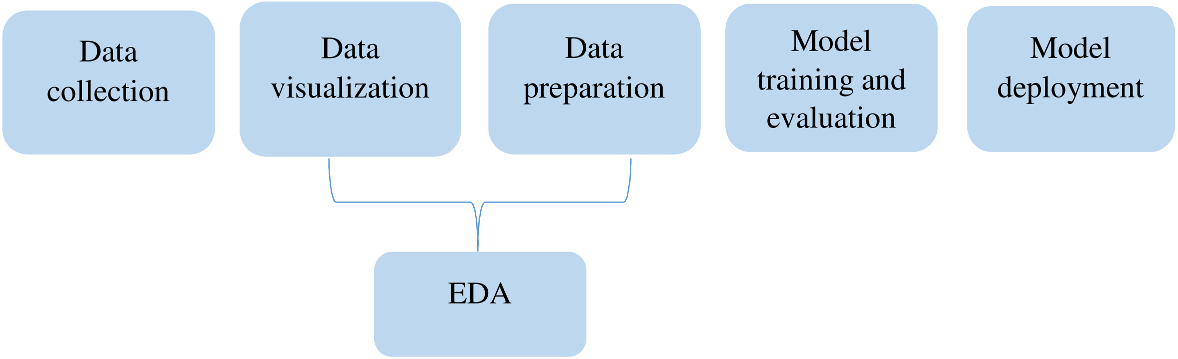


Figure 4: Predictive modeling process

These steps are going to be discussed in depth in the following chapter.

Below is the Gantt chart that shows the project’s main tasks against a timescale.

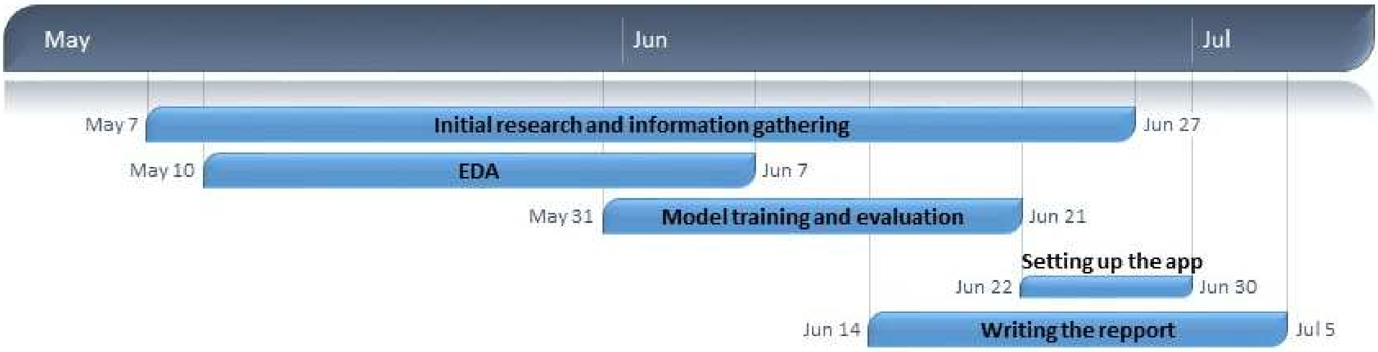


Figure 5: Gantt chart of the project

Chapter 2: Predictive Modeling: EDA and Machine Learning

I. Exploratory Data Analysis Overview

1. Data Analysis

Data analytics is the science of analyzing raw data in order to make conclusions about that information. It is a broad term that includes many diverse types of analyzing techniques. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve Decision Making for example, increase the overall efficiency of a business or system.., these techniques can reveal trends and patterns that would otherwise be lost in the mass of information.

Data analysis is defined as the systematic application of statistical and logical techniques to describe the scope of data, modularize the data structure, condense the data representation, illustrate via images, tables, and graphs, and evaluate statistical inclinations, probability data, and derive meaningful conclusions in the purpose of extracting useful information from this data and taking the decision based upon its analysis.

There is a common misconception that data analysis and data analytics are the same thing. The generally accepted distinction is that: [9]

- Data analytics is the broad field of using data and tools for Decision Making.
- Data analysis, a subset of data analytics, refers to specific actions that include a specific process.

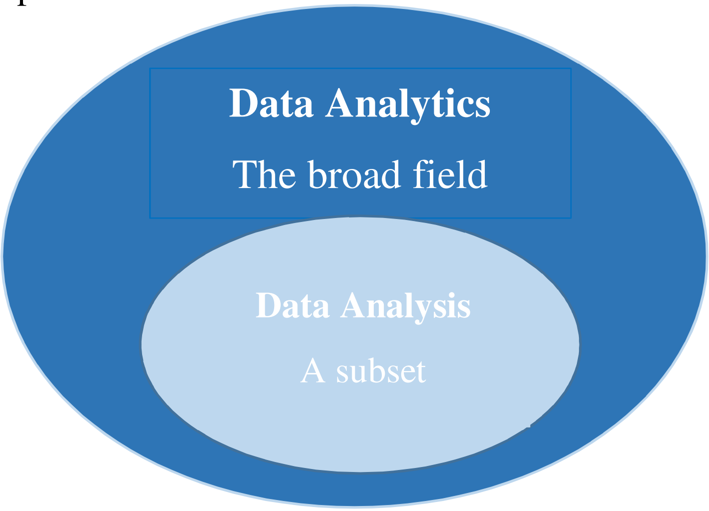


Figure 6: Data analytics and Data Analysis

An example of Data analysis in our day-to-day life is whenever we take any decision by recalling what has happened before or what will happen by choosing that particular one, this is an analysis of our past to base future decisions on it.

Data analysis skills and knowledge has become a necessity for making well-informed and efficient business decisions – as technology is fast changing how businesses operate – by helping the decision makers to understand problems facing the company and to explore data in meaningful ways, they have now a significant impact on how businesses are run.

There are several Data analysis tools with the purpose of processing and manipulating data, analyzing the relationships and correlations between datasets and identifying patterns and trends for interpretation:

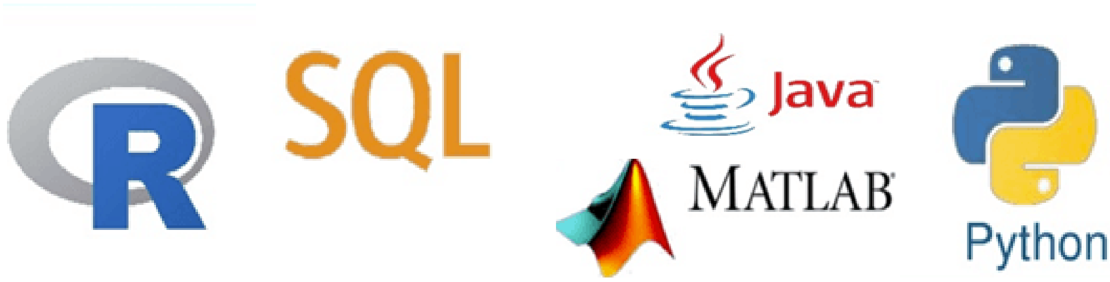


Figure 7: Data analysis tools

- SQL: Structured Query Language is a programming language used to access, read, manipulate, and analyze the data stored in a relational database and generate useful insights.
- R: a programming language that was designed to work with data at all stages of the data analysis process.
- Matlab: programming platform designed to access and analyze the data from a wide variety of sources.
- Java: a programming language that can be usable in a number of processes in the field of data science and throughout data analysis.
- Python: programming language with libraries that work perfectly for every stage of the data analysis process.

There are several types of Data Analysis techniques, the major ones are: [10]

- Statistical Analysis: the technique of performing several statistical operations to quantify the data and apply statistical analysis, there are two categories of this type of Analysis - Descriptive Analysis-analysis of the complete data or a sample of summarized numerical data, and Inferential Analysis-analysis of a sample from the complete data.
- Diagnostic Analysis: considered step further to statistical analysis, is referred to as "root cause" analysis as it includes processes like data discovery, mining…
- Prescriptive Analysis: suggests various courses of action and outlines the potential implications that could be reached after predictive analysis.
- Predictive Analysis: is the area of data analytics focused on interpreting existing data in order to make informed predictions about future events. It includes a variety of statistics techniques.

There are two popular data analysis approaches: [11]

- Classical
- Exploratory (EDA)

These two approaches are similar in that they all start with a general problem and all form conclusions. The difference is the sequence and focus of the intermediate steps:

In classical analysis, the data collection is the first step, then the imposition of a model (normality, linearity, etc.), the analysis, estimation, and testing that follows are focused on the parameters of that model.

For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate.

2. Predictive Analysis

Predictive Analysis is among the most useful applications of data analysis, Using it allows decision makers to predict upcoming challenges, identify opportunities for growth, and optimize their internal operations.

There isn't a single way to do predictive analytics, based on the goal, different methods provide the best results.

It includes a variety of statistics techniques: [12]

- Data mining: looking for patterns and relationships in large stores of data.
- Text analytics: deriving analysis-friendly structured data from unstructured text.
- Predictive modeling: creating and adjusting a statistical model to predict future outcomes.

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

3. Exploratory Data Analysis

EDA is an approach of analyzing datasets to summarize their main characteristics, it is how we get meaningful information from data often using statistical graphics and other data visualization method. It applies a variety of techniques to maximize specific insights into a dataset, reveal an underlying structure, extract significant variables, test hypotheses and assumptions, develop models, and determine best parameters for future estimations. [8]

The EDA approach is precisely that - an approach - not a set of techniques, but an attitude, a philosophy about how a data analysis should be carried out. [11]

It is the second step after problem understanding/data collection in CRISP methodology and it is essential for a well-defined and structured data analysis project, it should be performed before any statistical or machine learning modeling phase.

The two main pillars of EDA are: [8]

- Data preparation: cleaning the dataset, deleting non-relevant datasets, transforming the data, and dividing the data into required chunks for analysis.
- Data visualization and exploration: understanding the data, summarizing its characteristics, and visualizing it.

a. Data preparation

Data preparation - often referred to as "pre-processing"- is the act of manipulating raw data into a form that can readily and accurately be analyzed. To achieve the final stage of preparation, the data must be cleansed, formatted, and transformed into something digestible by analytics tools and algorithms.

Its purpose is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

Data can be gathered from any number of sources, and with that comes the possibility that it is not complete or fully accurate. To ensure that data is high quality, and therefore useful, it needs to be pre-processed before being used in a model. Otherwise, it will be a process of Garbage in, garbage out (GIGO); as the quality of data provided – the input - determines the quality of the output.

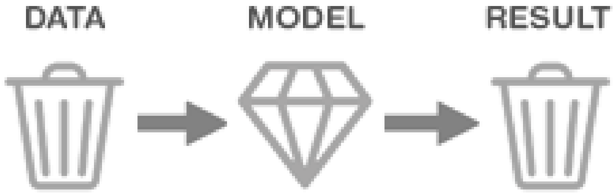


Figure 8: Garbage In Garbage Out

There are two main data types usually found in datasets:

- Quantitative data: measures of values or counts and are expressed as numbers: integer and real-valued floating point values.

- Qualitative data: categorical variables- measures of type -such as nominal, ordinal and Boolean.

The process for getting data ready for a machine learning algorithm can be summarized in three tasks: [8]

i. Select Data

Selecting the subset of all available data, there are some questions that would help get through this process:

- What is the extent of the available data? For example through time, database tables, connected systems.
- What data is not available that should have been? For example data that is not recorded or cannot be recorded. Deriving or simulating this data might be needed.
- What data that is not needed to address the problem? Excluding data is almost always easier than including data. Note down which data should be excluded and why.

ii. Preprocess Data

This preprocessing step is about getting the selected data into a form that can be used for analysis. Three common data preprocessing steps are formatting, cleaning and sampling:

- Formatting: Selected data may not be in a format that is suitable for modeling.
- Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data needed to address the problem, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely, but might as well be preserved treated.

- Sampling: There may be far more selected data available than what is needed to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. Taking a smaller representative sample of the selected data may be much faster for exploring and prototyping solutions before considering the whole dataset.
- It is very likely that the machine learning tools used on the data will influence the required preprocessing, revisiting this step might be needed.

Missing values are one of the most common problems encountered while preparing data for machine learning, the reason for the missing values might be human errors, interruptions in the data flow, privacy concerns, and so on… The simplest solution to the missing values is to drop the rows or the entire column.

However, other techniques can be used, like:

- Numerical Imputation: Imputation is a more preferable option rather than dropping because it preserves the data size, considering for example a possible default value of missing values in the column like ‘0’because it is likely that the NAN rows correspond to 0.
- Categorical Imputation: Replacing the missing values with the maximum occurred value in a column is a good option for handling categorical columns, or imputing a category like “Other” if the values in the column are distributed uniformly.

iii. **Transform Data**

The final step is to transform the data, three common data transformations are scaling, attribute decompositions and attribute aggregations. This step is also referred to as feature engineering, and consists of:

- Scaling: The preprocessed data may contain numerical attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume.

- Categorical data transformation: Machine learning models require all input and output variables to be numeric, this means if the data contains categorical attributes must be encoded to numbers.
- Decomposition: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant to the problem being solved.
- Aggregation: There may be features that can be aggregated into a single feature that would be more meaningful to the problem. For example, there may be a data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded.

Additional tasks can be included in Data preparation, such as: [13]

- Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. Feature selection techniques are used for several reasons.

- Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data.

b. Data Visualization

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed.

The main goal of data visualization is effectively, efficiently, elegantly, accurately as well as meaningfully communicating information. It fulfills its objectives only if it encodes the given input in such a manner that our eyes can recognize and our brain can comprehend.

There most popular Data Visualization types are: [8]

- Univariate Visualization: provides summary statistics for each field in the raw dataset summary only on one variable, histograms for example

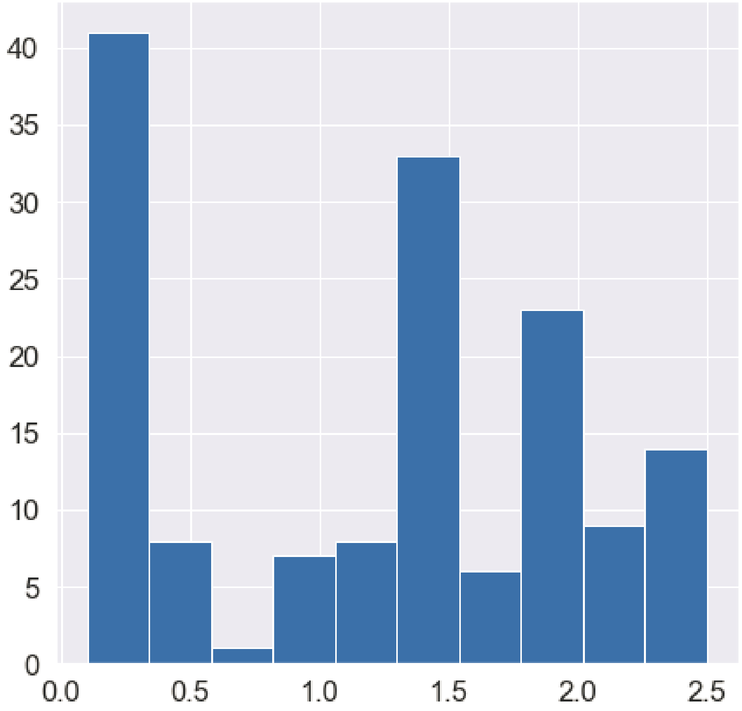


Figure 9: Histogram example

- Bivariate analysis: is performed to find the relationship between each variable in the dataset and the target variable.

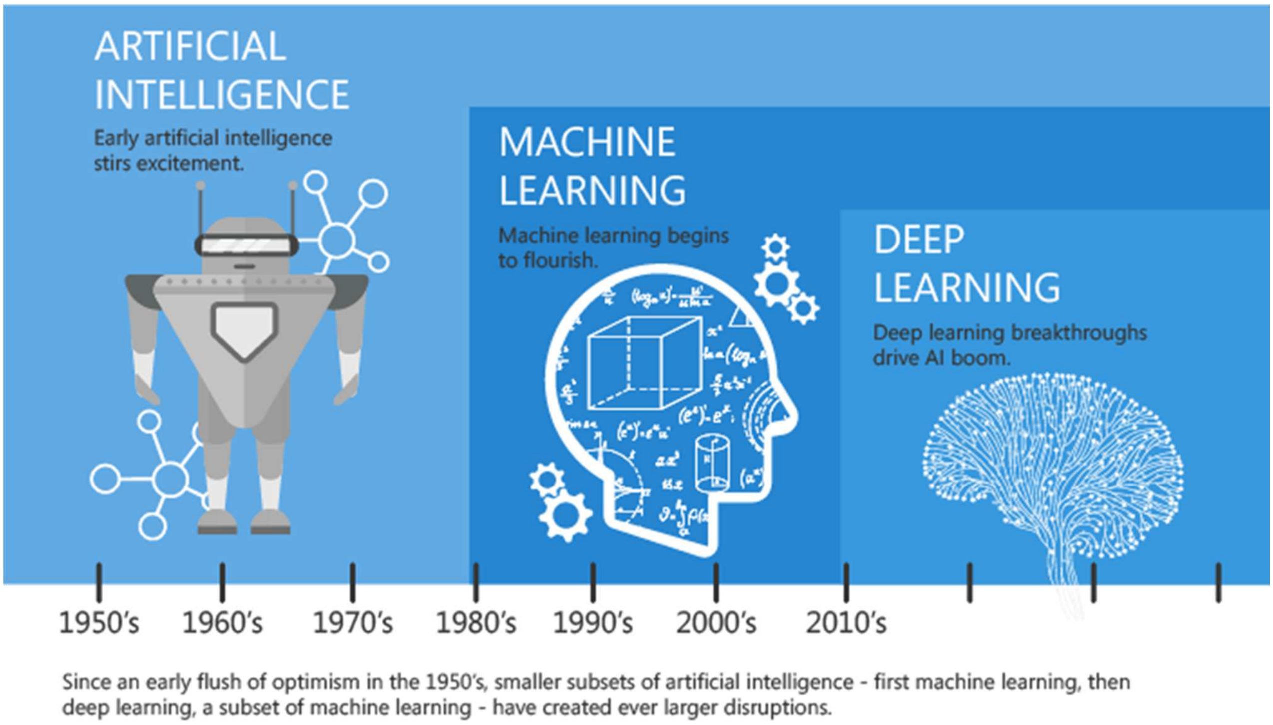
II. Machine Learning Overview

1. Artificial Intelligence

AI is the simulation of human intelligence processes by machines, especially computer systems.

AI technologies are categorized by their capacity to mimic human characteristics, the technology they use to do this, their real-world applications, and the theory of mind, using these characteristics for reference, all artificial intelligence systems - real and hypothetical - fall into one of three types: [14]

- Artificial narrow intelligence (ANI), which has a narrow range of abilities
- Artificial general intelligence (AGI), which is on par with human capabilities
- Artificial superintelligence (ASI), which is more capable than a human.


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

Figure 10: Quick history of AI

It is a common misconception that AI and ML are the same thing, but Machine learning is a subset of AI that provides systems the ability to automatically learn and improve from experience without being “explicitly programmed” as Arthur Samuel defined it. AI is a broad term - more of an umbrella term - that simply means making computers act intelligently.

2. Machine Learning : Models and Algorithms

ML is a type of AI. I t is the study of computer algorithms that improve automatically through experience and by the use of data. An ML model is a file that has been trained to recognize certain types of patterns.

Tom Mitchell – another Computer Scientist and machine learning pioneer – provided a more formal definition: "**A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.**"

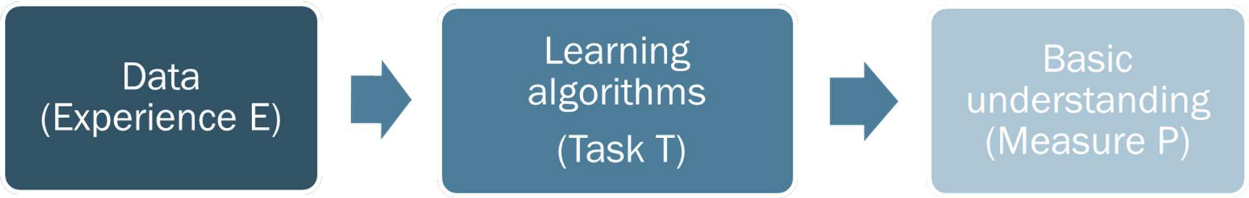


Figure 11: Machine learning as defined by Tom Mitchell

Machine learning algorithms are mainly divided into four categories: [15]

- Supervised learning: Supervised learning is a machine learning method in which models are trained using labeled data. In supervised learning, models need to find the mapping function to map the input variable (X) with the output variable (Y), supervision is needed to train the model .It can be used for two types of problems: Classification and Regression. A classification problem is when the output variable is a category, and a regression problem is when the output variable is a real value.
- Unsupervised learning: machine learning method in which patterns inferred from the unlabeled input data. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning does not

need any supervision. Instead, it finds patterns from the data by its own. Unsupervised learning can be used for two types of problems: Clustering and Association. A clustering problem is where the inherent groupings in the data are discovered, and an association rule learning problem is where rules that describe large portions of data are discovered.

- Semi-supervised learning: In one task, a machine learning model that automatically uses a large amount of unlabeled data to assist learning directly of a small amount of labeled data.
- Reinforcement learning: It is an area of machine learning concerned with how agents ought to take actions in an environment to maximize some notion of cumulative reward. The difference between reinforcement learning and supervised learning is the teacher signal. The reinforcement signal provided by the environment in reinforcement learning is used to evaluate the action (scalar signal) rather than telling the learning system how to perform correct actions.

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets, labels or categories. A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more of a set of classes.

There is a lot of classification algorithms, but it is not possible to conclude which one is superior to other as it depends on the application and nature of available data set.

a. Naïve Bayes Classifier

Naive Bayes is a statistical classification technique based on Bayes Theorem with an assumption of independence among features, it is one of the simplest supervised learning algorithms and is a reliable algorithm that has high accuracy and speed on large datasets.
The classifier assumes that the effect of a particular feature in a dataset is independent of other features, even if the features are interdependent, they are still considered independently. This assumption simplifies computation, and that's why it is considered naïve, this assumption is called class conditional independence. [16]

The Bayes theorem upon which this classifier is based, provides a way of calculating the posterior probability, P(c|x) – which is the probability of an event occurring given that another event has occurred- , from P(c), P(x), and P(x|c). [17]

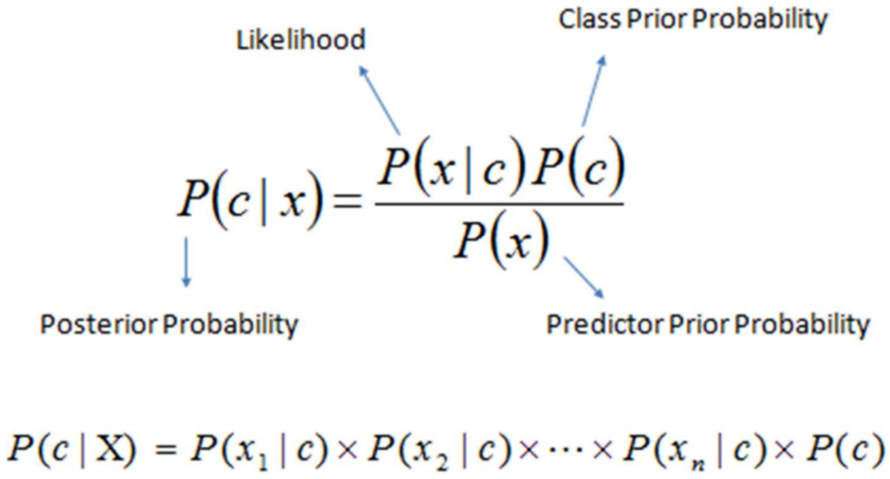

Likelihood
Class Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$

Figure 12: The Bayes theorem

Where:

- P(c|x) is the posterior probability of class (target/output) given feature.
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of feature given class
- P(x) is the prior probability of feature [17]

A Naive Bayes classifier calculates the probability of an event in the following steps:

- Step 1: Calculate the prior probability for given class labels
- Step 2: Find Likelihood probability with each attribute for each class
- Step 3: Put these value in Bayes Formula and calculate posterior probability.
- Step 4: See which class has a higher probability, given the input belongs to the higher probability class. [16]

Advantages and disadvantages of Naïve Bayes classification: [18]

- Advantages:

  - It is easy and fast to predict class of test data set. It also perform well in multi class prediction.
  - When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression.
  - It perform well in case of categorical input variables compared to numerical variables; for numerical variable, normal distribution is assumed bell curve, which is a strong assumption.
- Disadvantages:
  - If categorical variable has a category which was not observed in training dataset, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use a smoothing technique like Laplace estimation.
  - On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously.
  - In real life, it is almost impossible that the features are completely independent.

b. Random Forest Classifier

Random forest is a supervised ensemble learning algorithm, it builds multiple decision trees and merges them together to get a more accurate and stable prediction, the output is the class that is the mode of the classification or mean/average prediction – in a regression problem - of the individual trees.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label – the decision taken after computing all attributes.

A Random Forest Classification process is: [19]

- Step 1: The selection of random samples from a given dataset.
- Step 2: The algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- Step 3: voting will be performed for every predicted result.
- Step 4: select the most voted prediction result as the final prediction result.

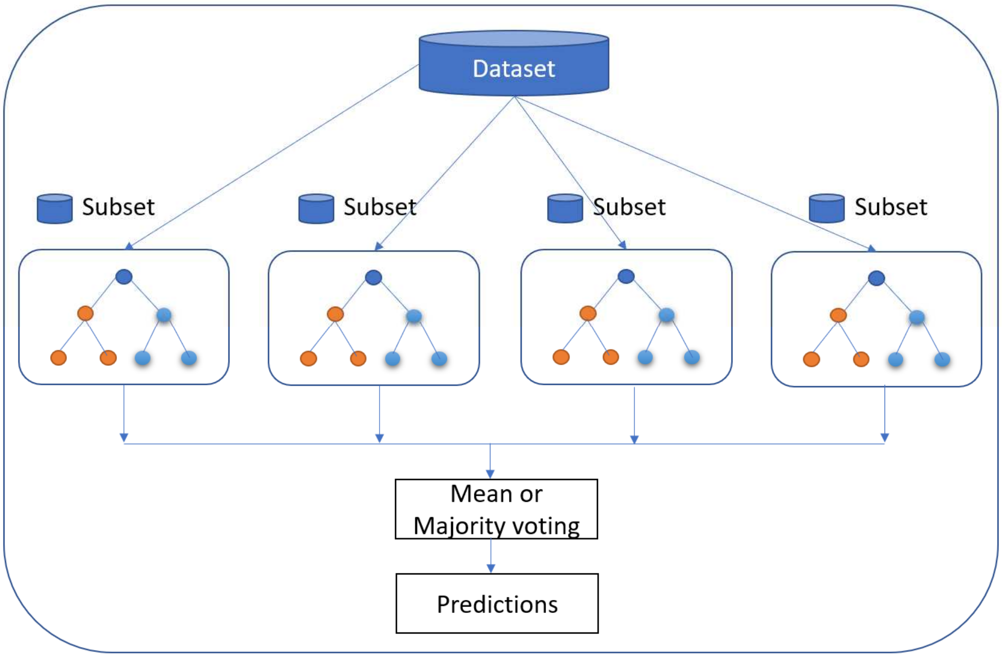


Figure 13: Random Forest classification Process

Advantages and disadvantages of a Random Forest Classification:

- Advantages:
  - Random forests work well for a large range of data items than a single decision tree does and has less variance then single decision tree.
  - Random forests are very flexible and possess very high accuracy.
  - They good accuracy even a large proportion of the data is missing.
- Disadvantages:
  - Complexity is the main disadvantage of Random forest algorithms.
  - Construction of Random forests are much harder and time-consuming than decision trees.
  - More computational resources are required to implement Random Forest algorithm.
  - It is less intuitive in case when we have a large collection of decision trees. [20]

c. Multi-Layer Perceptron Classifier

ANNs, or simply called neural networks – NNs, are computing systems vaguely inspired by the biological neural networks that constitute human brain, it is based on a collection of connected units or nodes called artificial neurons. They have the unique ability to derive meaning from complex and imprecise data, their power comes from, their ability to learn the representation in the training data and how to best relate it to the output variable. There are six types of neural networks, but one of the most popular ones is a feedforward neural network, it sends data in one direction only: from input nodes, through hidden nodes - if any exist - to the output nodes. [21]

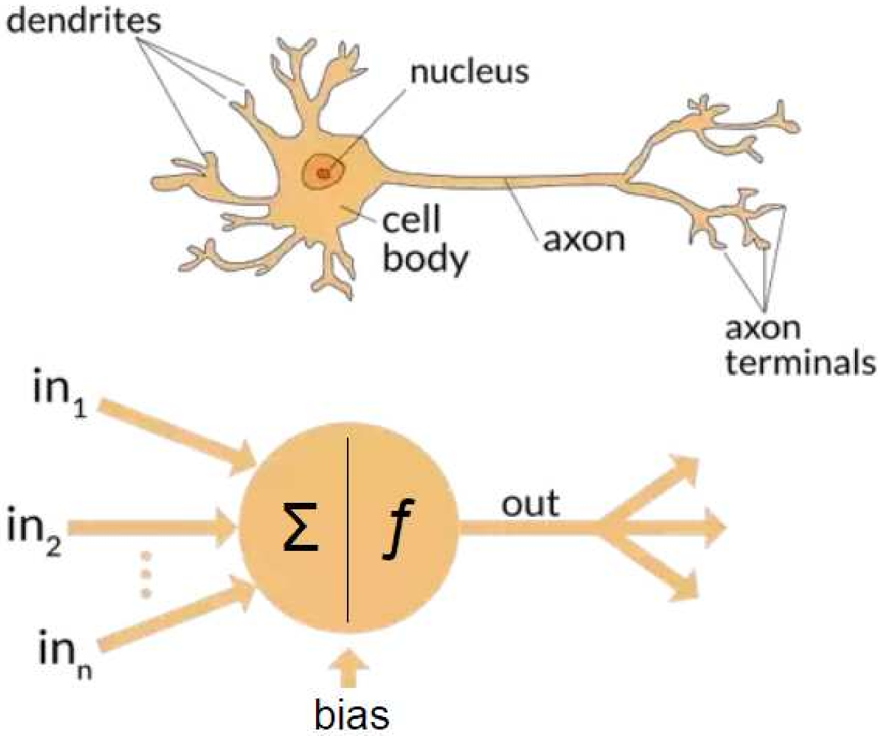


Figure 14: Artificial and biological neurons [22]

ANN components are:

- Neurons: ANNs are composed of artificial neurons – perceptrons, each artificial neuron has inputs and produces a single output. The inputs can be the feature values of a sample of external data, such as images or documents, or they can be the outputs of other neurons.

- Weights and connections: consists of connections, each connection providing the output of one neuron as an input to another neuron. Each connection is assigned a weight that represents its relative importance, a given neuron can have multiple input and output connections.
- Propagation function: computes the input to a neuron from the outputs of its predecessor neurons and their connections as a weighted sum for computation, each neuron considers weights and bias - an additional input of 1. Then, the combination function uses the weight and the bias to give an output – a modified input.

ANNs workflow:

- Information is fed into the input layer - a row of perceptrons that are not connected to each other, which transfers it to the hidden layer
- The connections between the two layers assign weights to each input randomly
- A bias added to every input after weights are multiplied with them individually
- The weighted sum is transferred to the activation function
- The activation function determines which nodes it should fire for feature extraction
- The model applies an application function to the output layer to deliver the output
- Weights are adjusted, and the output is back-propagated to minimize error.

Multi-layer Perceptron classifier is a class of feedforward artificial neural network – ANN - that utilizes a supervised learning technique called backpropagation for training to perform the task of classification.

An MLP classifier consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP layers: [23]

- The input layer: gathers the data from the outside world.

- The hidden layer performs all the back-end tasks of calculation, networks can even have zero hidden layers or multiple ,
- The output layer transmits the final result of the hidden layer's calculation.

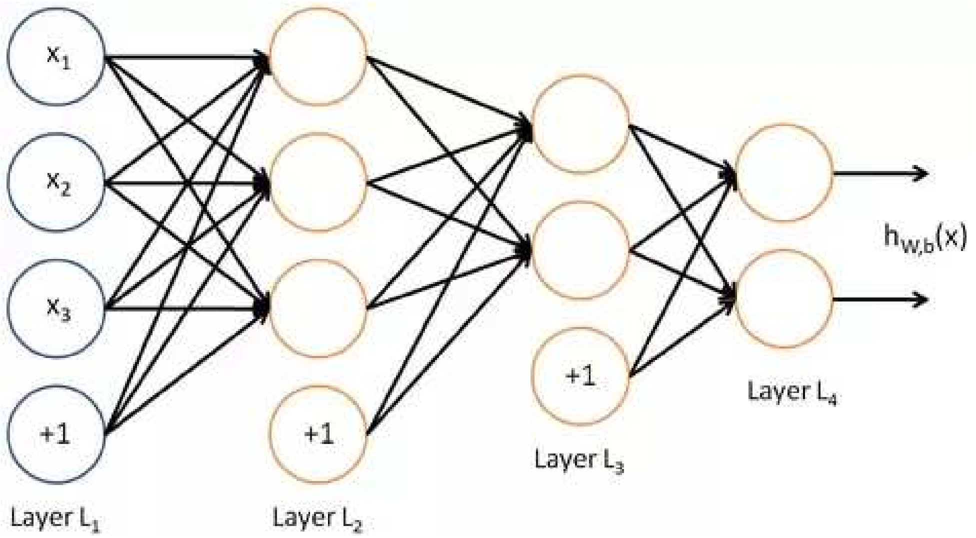


Figure 15: MLP Classifier nodes, layers and bias

Advantages and disadvantages of MLP classification: [24]

- Advantages :
  - Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
  - MLP classifiers form the required decision function directly via training.
  - Neural networks in general work best with more data points.
- Disadvantages:
  - Classification with MLP is computationally very expensive and time consuming to train; too many parameters because it is fully connected- parameter number = width x depth x height.
  - Neural networks depend a lot on training data. This leads to the problem of over-fitting and generalization often.

3. Machine Learning: Models Evaluation

After model building and training, it is a necessary and a logical step to evaluate them to answer questions like: How well is the model doing? Is it a useful one? …

To properly evaluate the model, it is important to not train it on new data to prevent the likelihood of overfitting to the training set. A typical train/test split would be to use 80% of the data for training and 20% of the data for testing. [13]

a. Metrics of performance

i. Classification

In classification, four types of outcomes could occur: [25]

- True positives are when the predicted observation belongs to a class and it actually does belong to that class.
- True negatives are when the predicted observation does not belong to a class and it actually does not belong to that class.
- False positives occur when the predicted observation belongs to a class when in reality it does not.
- False negatives occur when the predicted an observation does not belong to a class when in fact it does.

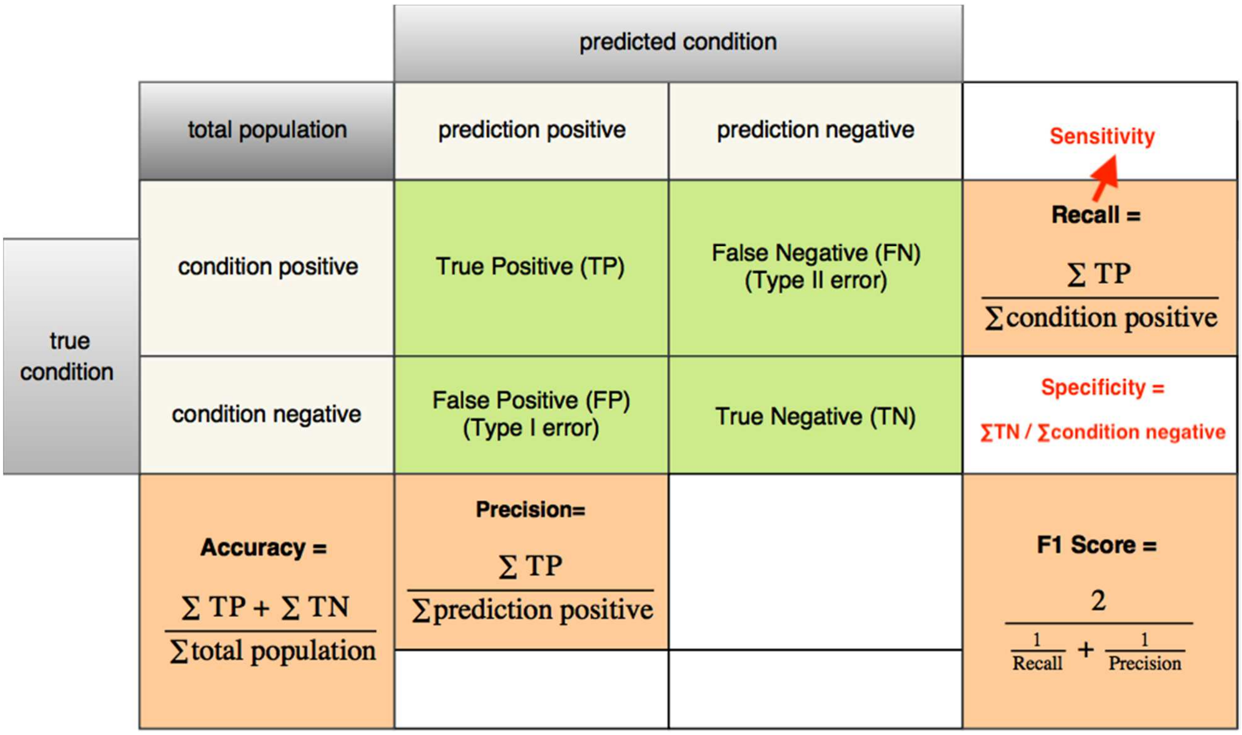
These four outcomes are often plotted on a confusion matrix.

ii. Accuracy, F1 score and recall in classification

There are four main metrics used to evaluate a classification model: [25]

- Accuracy: defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.
- Precision: defined as the fraction of relevant examples true positives among all of the examples which were predicted to belong in a certain class.

- Recall: defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.
- F1 score: combining the precision and recall of the model, it is defined as the harmonic mean of the model's precision and recall



		predicted condition		
	total population	prediction positive	prediction negative	Sensitivity
true condition	condition positive	True Positive (TP)	False Negative (FN) (Type II error)	Recall = $\frac{\Sigma\,\mathrm{TP}}{\Sigma\,\text{condition positive}}$
	condition negative	False Positive (FP) (Type I error)	True Negative (TN)	Specificity = ΣTN / Σcondition negative
	Accuracy = $\frac{\Sigma\,\mathrm{TP} + \Sigma\,\mathrm{TN}}{\Sigma\,\text{total population}}$	Precision= $\frac{\Sigma\,\mathrm{TP}}{\Sigma\,\text{prediction positive}}$		F1 Score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

Figure 16: Classification: metrics of performance

iii. Bias – variance tradeoff: Learning curves

In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.

Four main concepts to retain: [26]

- Bias: High bias occurs when the learning algorithm is not taking into account all the relevant information, becoming unable to capture the model’s richness and complexity.

- Variance: High variance happens when the model is too complex and doesn't represent the simpler real patterns existing in the data
- Underfitting: When the algorithm is not able to model either training data or new data, consistently obtaining high error values that don't decrease over time.
- Overfitting: The algorithm captures well the training data, but it performs poorly on new data, so it's not able to generalize.

Learning curves are plots that compares the performance of a model on training and testing data over a varying number of training instances.

Types of learning curves: [27]

- Bad Learning Curve: High Bias, underfit :
  - When training and testing errors converge and are high
  - No matter how much data is fed to the model, the model cannot represent the underlying relationship and has high systematic errors
  - Poor fit poor generalization

- Bad Learning Curve: High Variance, overfit:
  - When there is a large gap between the errors
  - Require data to improve
  - Can simplify the model with fewer or less complex features
- Ideal Learning curve: good fit
  - Model that generalizes to new data
  - Testing and training learning curves converge at similar values

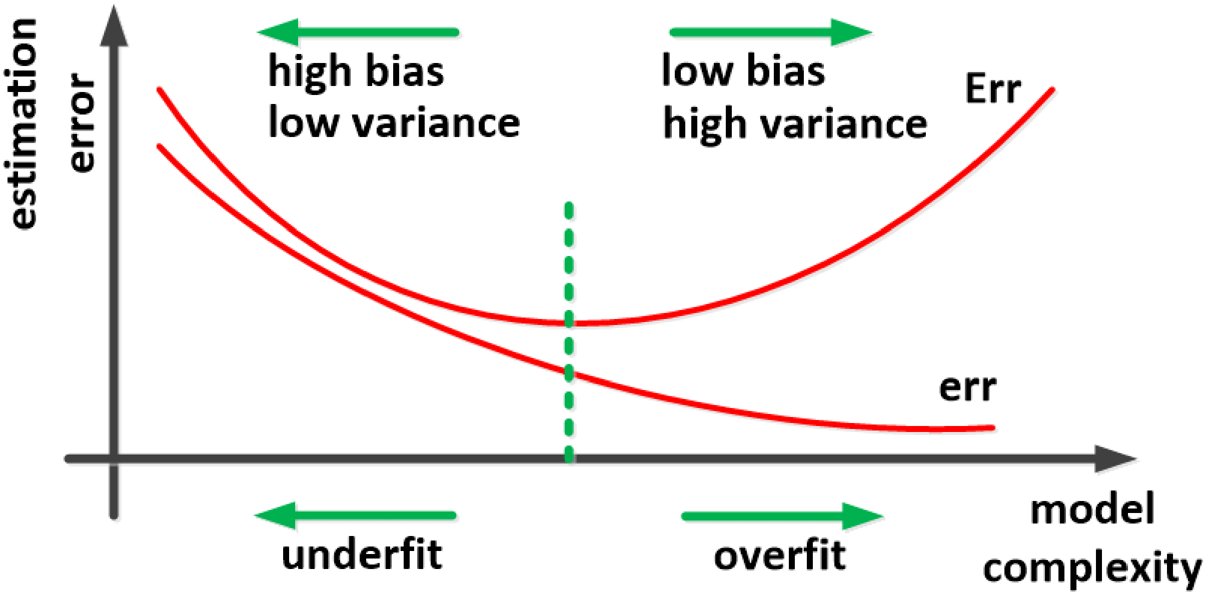


Figure 17: Overfit and underfit learning curve [28]

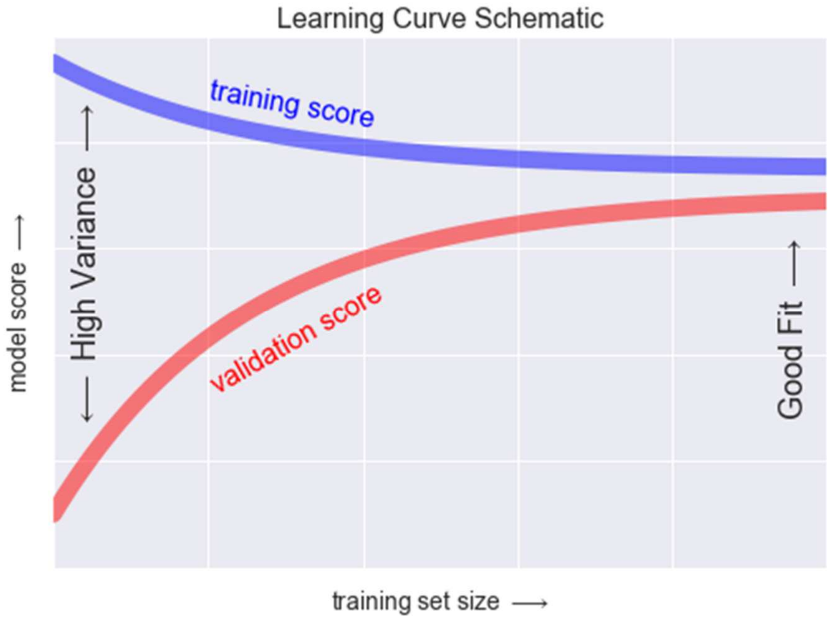


Figure 18: Good fit learning curve

b. Cross Validation

Cross-validation is a model evaluation technique. It is a resampling procedure used to evaluate machine learning models on a limited data sample.

K-Folds Cross Validation is a cross validation technique primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data, its procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

The general procedure consists of: [29]

- Shuffling the dataset randomly
- Splitting it into k groups
- For each unique group:
  - A group is hold as a testing set
  - The remaining groups are the training sets
  - The model is fit on the training set and evaluated on the test set
  - The evaluation score is retained and the model is discarded
- The performance of the model is summarized using each sample of model evaluation scores

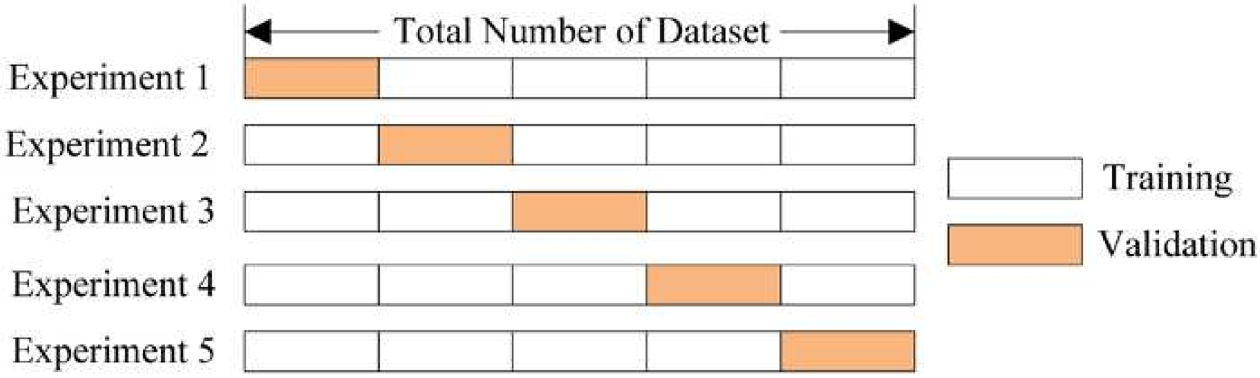


Figure 19: K-Fold Cross Validation

Generally, a stratified 10-fold cross-validation is the best practice for classification. [13]

4. Machine Learning: Hyperparameters tuning

Hyperparameter tuning, or hyperparameter optimization is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

The trade-off between bias and variance as mentioned before is determined by the complexity of the model and the amount of training data. The optimal hyperparameters help to avoid under-fitting -training and test error are both high, and over-fitting - training error is low but test error is high.

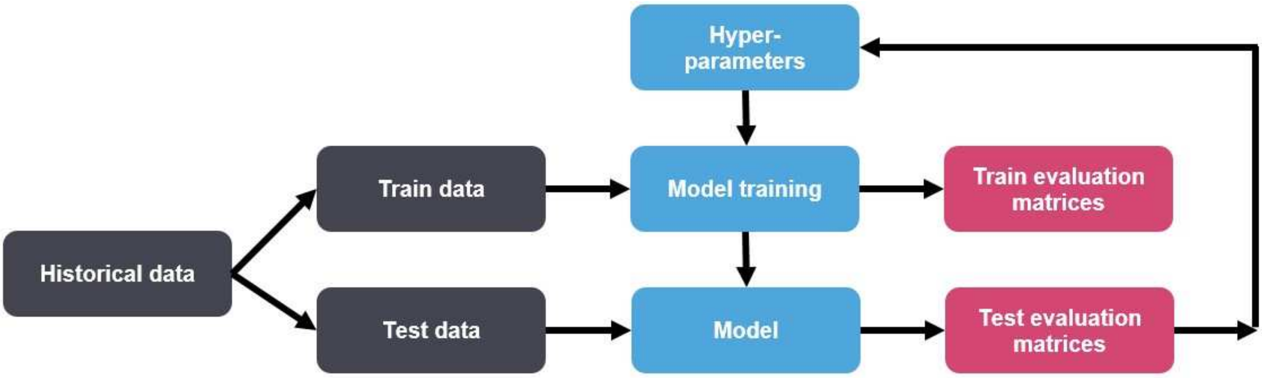


Figure 20: Model Evaluation mechanism [30]

A range of different optimization algorithms may be used, although two of the simplest and most common methods are:

- Random Search: defines a search space as a bounded domain of hyperparameter values and randomly samples points in that domain.
- Grid Search: defines a search space as a grid of hyperparameter values and evaluates every position in the grid

Chapter 3: Customer Churn Prediction Application

I. EDA and modeling

1. Used tools

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its notable use of significant indentation, its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. It is dynamically-typed and garbage-collected and supports multiple programming paradigms, including structured - particularly, procedural, object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python was introduced back in 1991 by its creator Guido van Rossum but it began to gain popularity only a couple of years ago. In 2020, it became the fourth most used language after JavaScript, HTML/CSS, and SQL, with 44, 1% of developers using it for API development, Artificial Intelligence, web development, Internet of Things….

As mentioned before, Python works well on every stage of data analysis and is widely used for it. It is the libraries that were designed for data science that are so helpful. Data mining, data processing and modeling along with data visualization are the 3 most popular ways of how Python is being used for data analysis.

I chose to use Python - over many other tools of Data Analysis, to perform EDA and modeling on the provided dataset because it is easy to learn, due to its clear syntax and readability and because of its wide range of libraries that can be used for each stage of data analysis and modeling.

a. Python libraries for EDA

i. Data Visualization

- Matplotlib: is the plotting library for Python that provides an object-oriented API for embedding plots into applications. It is a close resemblance to MATLAB embedded in Python programming language Matplotlib can depict a wide range of visualizations: histogram, bar plots, scatter plots... and with a bit of effort and tint of visualization capabilities, any visualization can be created, it also facilitates labels, grids, legends, and some more formatting entities with Matplotlib

- Seaborn: is defined in the official documentation as the data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics; Seaborn is an extension of Matplotlib with advanced features. While Matplotlib is used for basic plotting; bars, pies, lines, scatter plots and stuff, Seaborn provides a variety of visualization patterns with less complex and fewer syntax.

Figure 21: Data Visualization libraries

ii. Data Preprocessing

- Pandas: Pandas or Python Data Analysis Library is an open-source Python package that provides high-performance, easy-to-use data structures and data analysis tools for the labeled data in Python programming language. It is designed for quick and easy data manipulation, reading, aggregation, Indexing, renaming, sorting, merging dataset, handling missing data or NANs…. Making it a foundation library in learning Python for Data Analysis. Pandas take data in a CSV or TSV file or a SQL database and create a Python object with rows and columns called a dataframe. The dataframe is very similar to a table in statistical software, say Excel for example.

- NumPy: is a general-purpose array-processing package, one of the most fundamental packages in Python. It provides high-performance multidimensional array objects and tools to work with the arrays. NumPy is an efficient container of generic multi-dimensional data. It is used to process arrays that store values of the same datatype, as it facilitates math operations on arrays and their vectorization which significantly enhances performance and speeds up the execution time correspondingly.

- Sklearn.Preprocessing: it is a package - of the Sklearn library - that provides several common utility functions and transformer classes to prepare input data as a text for processing with machine learning algorithms.

Figure 22: Data preprocessing libraries and packages

b. Sklearn

Scikit Learn is a robust machine learning library for Python. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. It is one of the easiest and cleanest ML library.

I chose to use Sklearn because it is simple and easy to learn and offers a variety of tools to help pick the good –say correct - model, the used packages include:

- Preprocessing: as mentioned before, it is a preprocessing package that Sklearn provide even though it focuses primarily on modeling data not manipulating data.

- Naive_bayes: includes supervised learning methods based on applying Bayes' theorem with naive feature independence assumptions, an example of this methods id the GaussianNB classifier, which is used in classification and assumes that features follow a normal distribution.

- Neural_network: includes models based on neural networks, like the MLPClassifier, which is, stated in the official documentation, that it is not suitable for large datasets.

- Ensemble: a set of estimators that combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator, includes the RandomForestClassifier.

- Metrics: module includes score functions, performance metrics and pairwise metrics and distance computations, like accuracy-score…..

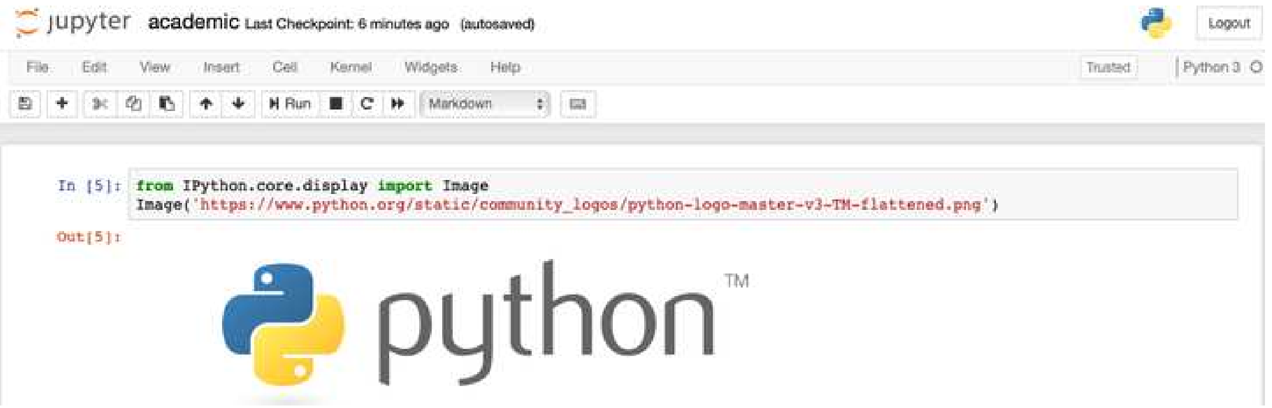
- Model-selection: includes methods of model validation like StratifiedKFold Cross Validation, learning curves, hyper parameter tuning like GridSearch and splitter functions like the train_test_split function.

c. IDE

The Jupyter Notebook is an open-source web application that allows users – usually data scientists, analysts… to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document. The name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R. It is in an easy-to-use, interactive environment that a way of working inside a virtual "notebook"

A Jupyter notebook has two components. First, the programming code or text is entered in rectangular "cells" in a front-end web page, the browser then passes the code to a back-end "kernel" which runs the code and returns the results. The kernels need not reside on the users' computers, as it is more of a virtual notebook.

Jupyter Notebooks can be used for all sorts of data science tasks including data cleaning and transformation, numerical simulation, exploratory data analysis, data visualization, statistical modeling, machine learning, deep learning… and is growing in popularity with data scientists in large part due to its flexibility, it gives them a way to combine code, images, plots, comments…, in alignment with the step of the data science process, which is why I chose to use it.



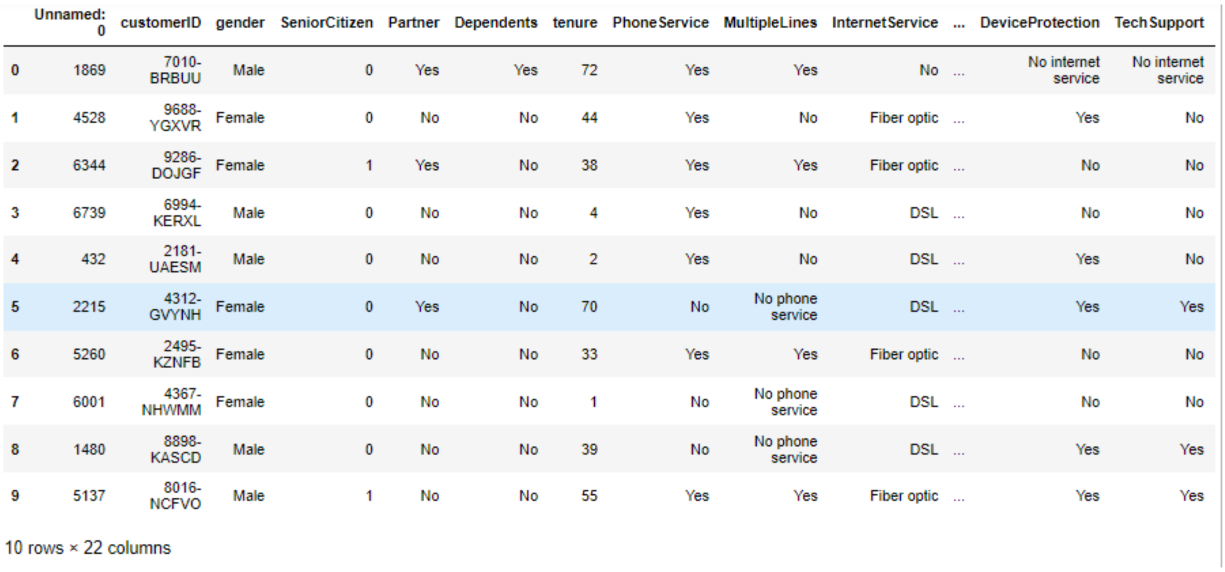
```
In [5]: from IPython.core.display import Image
        Image('https://www.python.org/static/community_logos/python-logo-master-v3-TM-flattened.png')
Out[5]:
```

Figure 23: Jupyter notebook interface

2. Performing EDA and modeling

a. Exploratory Data Analysis

To perform EDA techniques on the dataset provided by the telecom company I had to load the dataset to the notebook first, as well as all the needed Python libraries. I had to check the first 5 entries to get familiar with it and have a clear understanding of what it is like.



	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	...	DeviceProtection	TechSupport
0	1869	7010-BRBUU	Male	0	Yes	Yes	72	Yes	Yes	No	...	No internet service	No internet service
1	4528	9688-YGXVR	Female	0	No	No	44	Yes	No	Fiber optic	...	Yes	No
2	6344	9286-DOJGF	Female	1	Yes	No	38	Yes	Yes	Fiber optic	...	No	No
3	6739	6994-KERXL	Male	0	No	No	4	Yes	No	DSL	...	No	No
4	432	2181-UAESM	Male	0	No	No	2	Yes	No	DSL	...	Yes	No
5	2215	4312-GVYNH	Female	0	Yes	No	70	No	No phone service	DSL	...	Yes	Yes
6	5260	2495-KZNFB	Female	0	No	No	33	Yes	Yes	Fiber optic	...	No	No
7	6001	4367-NHWMM	Female	0	No	No	1	No	No phone service	DSL	...	No	No
8	1480	8898-KASCD	Male	0	No	No	39	No	No phone service	DSL	...	Yes	Yes
9	5137	8016-NCFVO	Male	1	No	No	55	Yes	Yes	Fiber optic	...	Yes	Yes

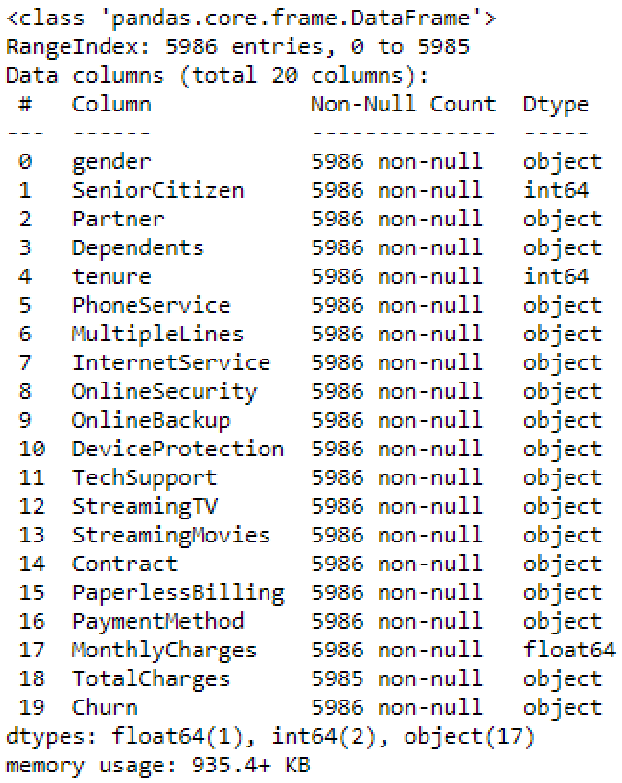
10 rows × 22 columns

Figure 24: First look at the dataset

The dataset contains 5986 rows and 22 columns where the last column ‘Churn’ is the label. The first column is unnamed and contains integer values so it is safe to assume that it is meant for indexing, the second column contains the customers’ IDs which makes it irrelevant for analysis, and these two columns need to get dropped.

The ‘SeniorCitizen’ column’s values are binary -1s and 0s, while it is actually in nature a categorical attribute, so in order to not get it mixed with numeric attributes the values should be replaced with ‘yes’ and ‘no’.

To make sure that every other attribute type in the dataset is valid, I had to check the dataset info.



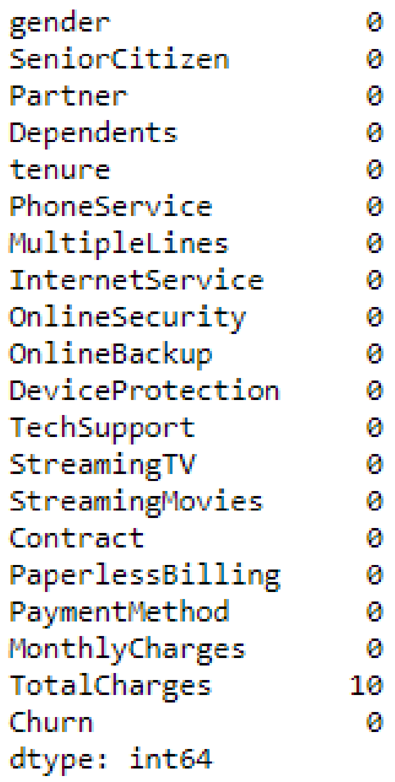
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5986 entries, 0 to 5985
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   gender            5986 non-null   object 
 1   SeniorCitizen     5986 non-null   int64  
 2   Partner           5986 non-null   object 
 3   Dependents        5986 non-null   object 
 4   tenure            5986 non-null   int64  
 5   PhoneService      5986 non-null   object 
 6   MultipleLines     5986 non-null   object 
 7   InternetService   5986 non-null   object 
 8   OnlineSecurity    5986 non-null   object 
 9   OnlineBackup      5986 non-null   object 
 10  DeviceProtection  5986 non-null   object 
 11  TechSupport       5986 non-null   object 
 12  StreamingTV       5986 non-null   object 
 13  StreamingMovies   5986 non-null   object 
 14  Contract          5986 non-null   object 
 15  PaperlessBilling  5986 non-null   object 
 16  PaymentMethod     5986 non-null   object 
 17  MonthlyCharges    5986 non-null   float64
 18  TotalCharges      5985 non-null   object 
 19  Churn             5986 non-null   object 
dtypes: float64(1), int64(2), object(17)
memory usage: 935.4+ KB
```

Figure 25: Dataset info

It turned out that, indeed, the ‘TotalCharges’ column type is ‘object64’ while it is a numeric attribute in nature, so it need to get converted to ‘float64’.
The column also seems to have a missing value so that should be treated.

Datasets with duplicated rows are not suitable for modeling as they bias the models, dropping them is a good practice. There was 16 duplicated rows in the dataset that have been removed.

Now for missing values, I had to check all the columns to see if there have been any empty values that have not been uncovered before.



```
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       10
Churn               0
dtype: int64
```

Figure 26: Count of empty values in the dataset

Additionally to the one found before, there are 9 other missing values in the ‘TotalCharges’ column and for that to be treated, I used the fact that the monthly charges are , in general, monthly charges multiplied by tenure.
So first I checked if these values are correlated in this sense:

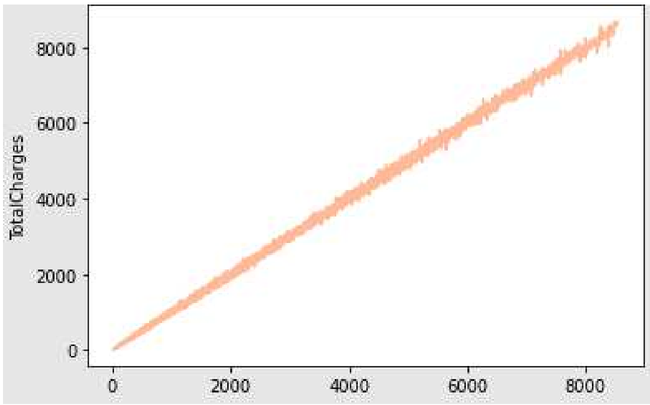
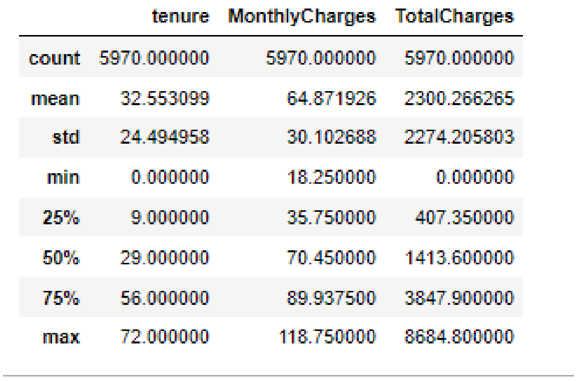


*Figure 27: TotalCharges by tenure*MonthlyCharges*

The plot shows that 'TotalCharges' column's values are proportional to the 'tenure' multiplied by 'MonthlyCharges' values, so the 'TotalCharges' column's empty values can be filled by the corresponding 'tenure' * 'MonthlyCharges'.

However, after printing the rows with the empty values, it turned out that they correspond to a '0' tenure value, so they can be filled with '0' as well.

To understand the dataset more, I checked the numerical attributes first.
I took a look at their description, which includes the 5 number summary, their count, mean and standard deviation.



	tenure	MonthlyCharges	TotalCharges
count	5970.000000	5970.000000	5970.000000
mean	32.553099	64.871926	2300.266265
std	24.494958	30.102688	2274.205803
min	0.000000	18.250000	0.000000
25%	9.000000	35.750000	407.350000
50%	29.000000	70.450000	1413.600000
75%	56.000000	89.937500	3847.900000
max	72.000000	118.750000	8684.800000

Figure 28: Numeric attributes description

For distribution analysis, the customers' distribution by these columns need to be visualized, the distribution skews need to be checked also.

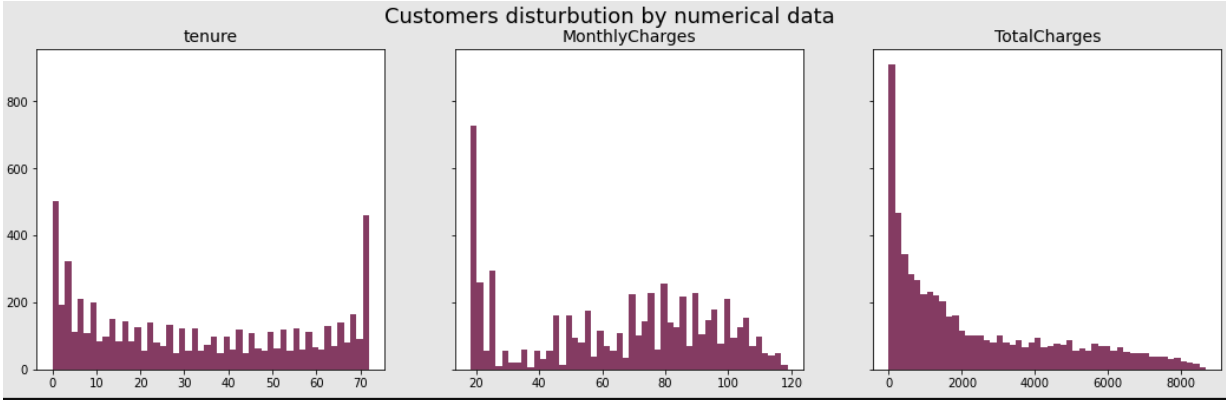
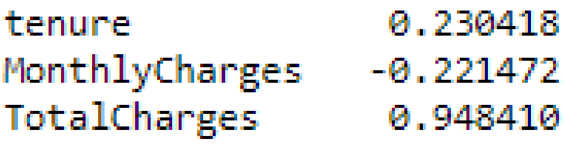


Figure 29: Customers distribution by numeric attributes

The ‘tenure’ column seems to be symmetrical, while the ‘MonthlyCharges’ two others are a little far from being symmetrical. However, the ‘tenure’ looks like an inverted bell curve, the ‘MonthlyCharges’ looks almost like a bell curve – minus the peaks on the left and the ‘TotalCharges’ seem to be right skewed.

The skews need to be analyzed in order to know the severity of the skewness.

Computed skews of the columns:

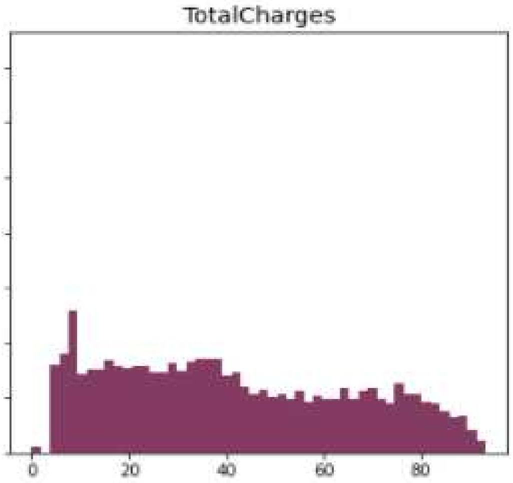


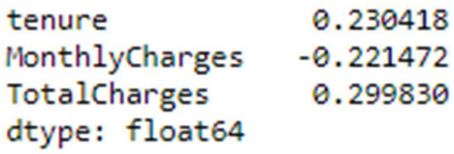
```
tenure             0.230418
MonthlyCharges    -0.221472
TotalCharges       0.948410
```

Figure 30: Numeric attributes distribution skews

According to the skewness values, 'tenure' and 'MonthlyCharges' ' plots are fairly symmetrical - the skewness is between -0.5 and +0.5 and quite close to 0, while ‘TotalCharges’'s plot is moderately right skewed - the skewness is between 0.5 and 1.

Since its skewness is not so severe, it might be corrected by a data transformation method, or just be left like that. I chose to apply a square root transformation on the ‘TotalCharges’ column.





```
tenure             0.230418
MonthlyCharges    -0.221472
TotalCharges       0.299830
dtype: float64
```

Figure 31: 'TotalCharges' skewness after transformation

Next, I checked the customer’ distribution by the categorical columns. First demographics related columns:

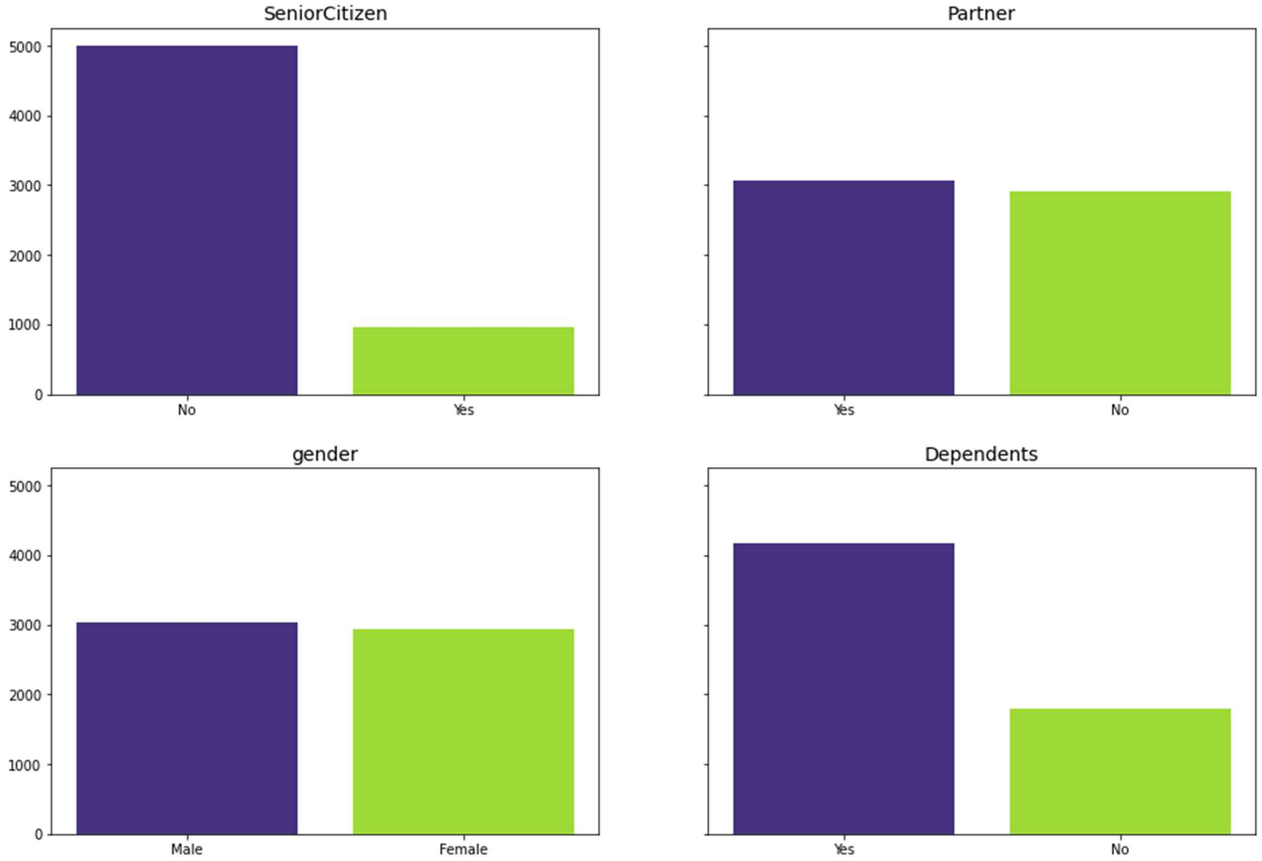


Figure 32: Customers distribution by demographics

Looking at the plots, the following conclusions can be drawn:

- Half of the customers in our data set are female and the other half are male.
- Same goes to partners, half the customers have partners.
- About 1/6 of customers are senior citizens.
- About 1/3 of customers have dependents.

Second, customers’ distribution by services:

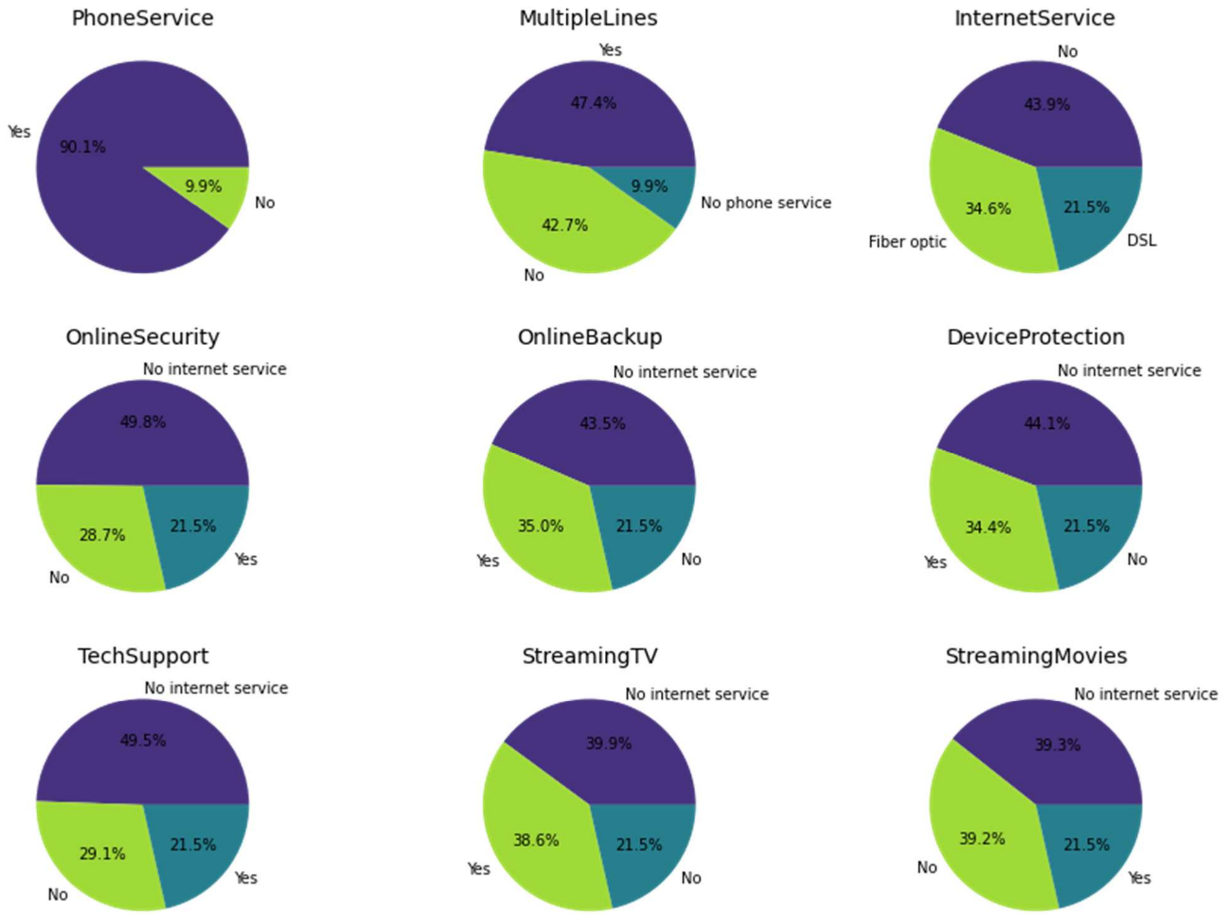


Figure 33: Customers distribution by services

Looking at the plots, we can conclude that:

- Most customers have a phone service and about half of them have multiple lines.
- Almost half of the customers don’t have internet or related services.

Finally, customers' distribution by payment preferences:

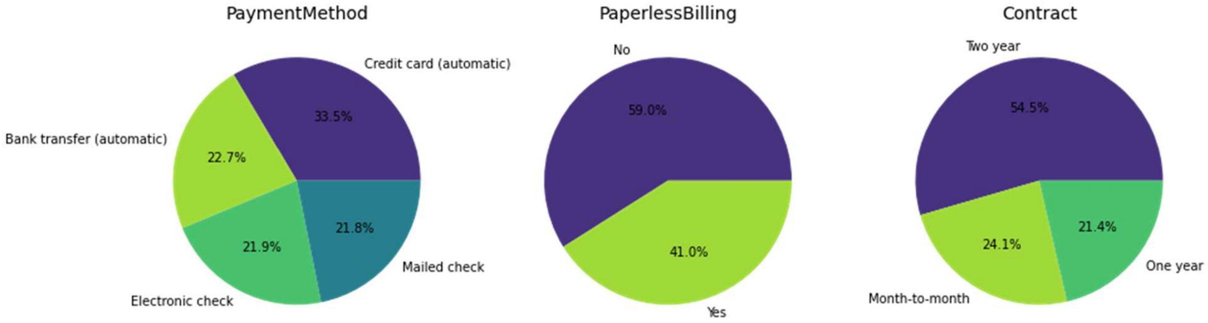


Figure 34: Customers' distribution by payment preferences

It is obvious that:

- Credit card payment is the most preferred method
- A little over half the customers prefer paperless billing and two-year type of contracts.

Now for the customers' distribution by the label – Churn:

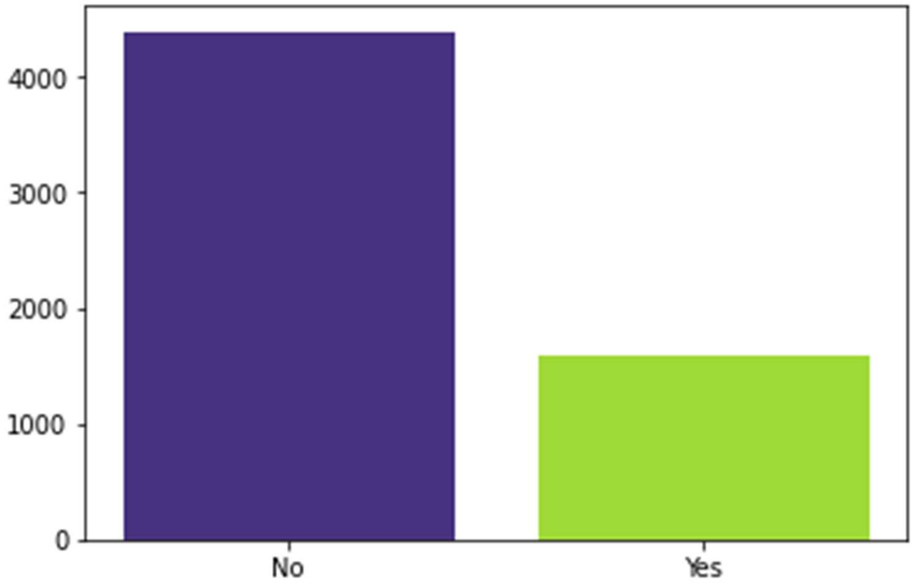


Figure 35: Customers churn distribution

- This churn rate is at approximately 25% making it moderately high.

The plot shows that about 75% of customers choose not to churn, leaving churn ratio at 20%, making the classes slightly imbalanced.

Churn correlation with the other attributes is to be checked next.

First by demographics:

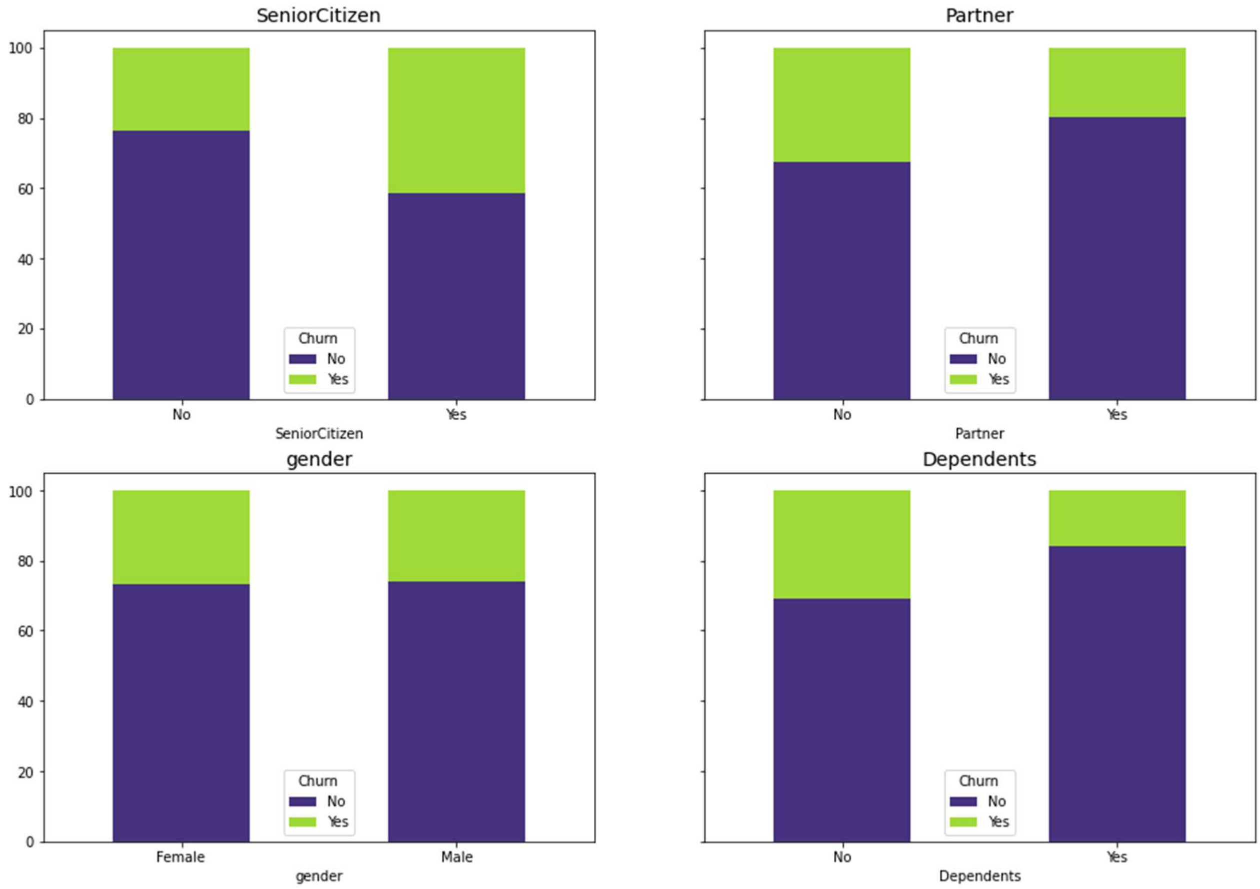


Figure 36: Churn distribution by demographics

Customers of both gender churn out the same, while those who are senior citizens and who have dependents and partners seem to choose to churn out the most.

Second, by services:

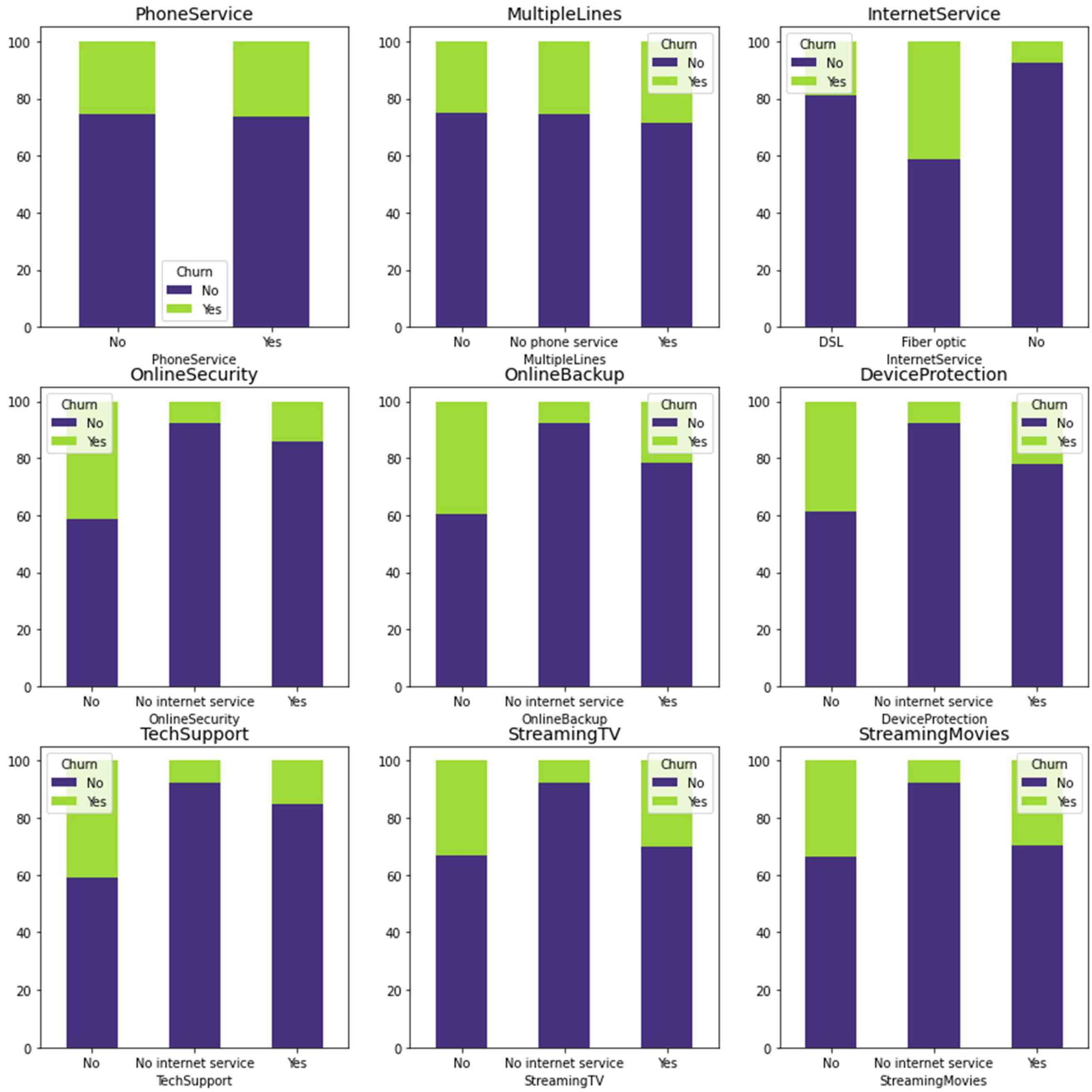


Figure 37: Churn distribution by service

- Customers with fiber optic option churn out the most.
- Customers with no internet service churn out the least.
- Customers with or without phone service have the same churn rate.
- Customers with no services are more likely to churn.

And finally, by payment preference:

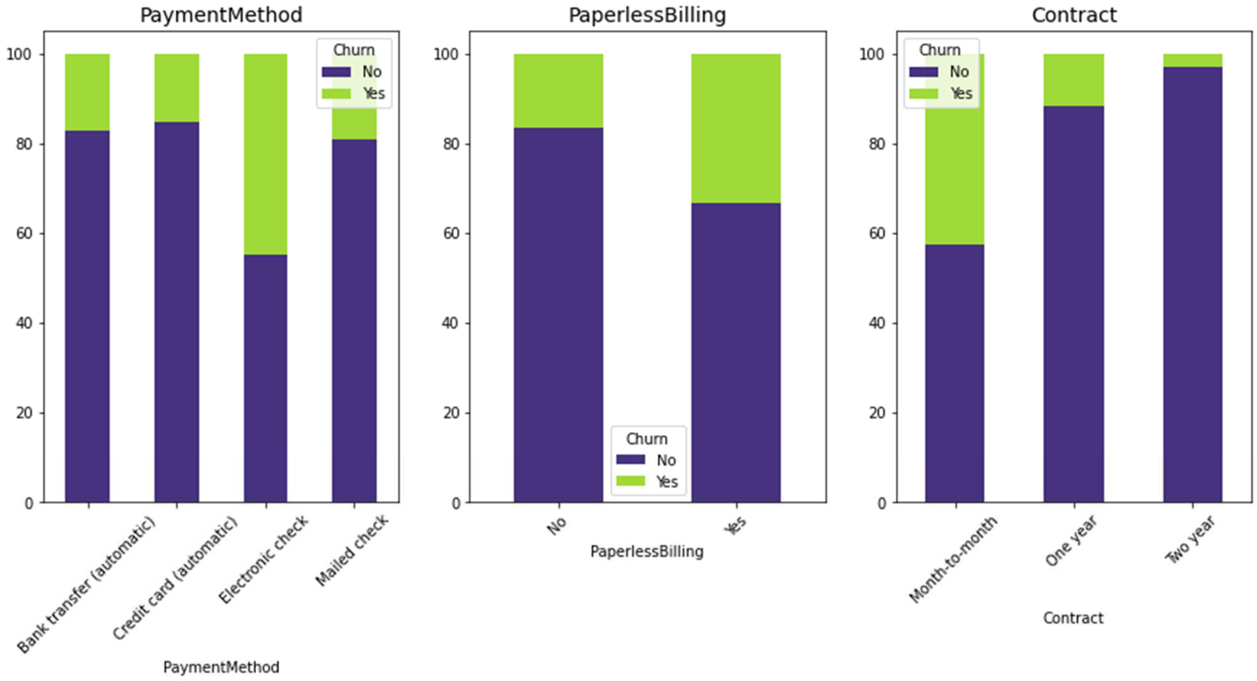


Figure 38: Churn distribution by payment preferences

- Users with month to month type of contract, paperless billing and electronic check payment method are churned out more often

Churn correlation to all of the values:

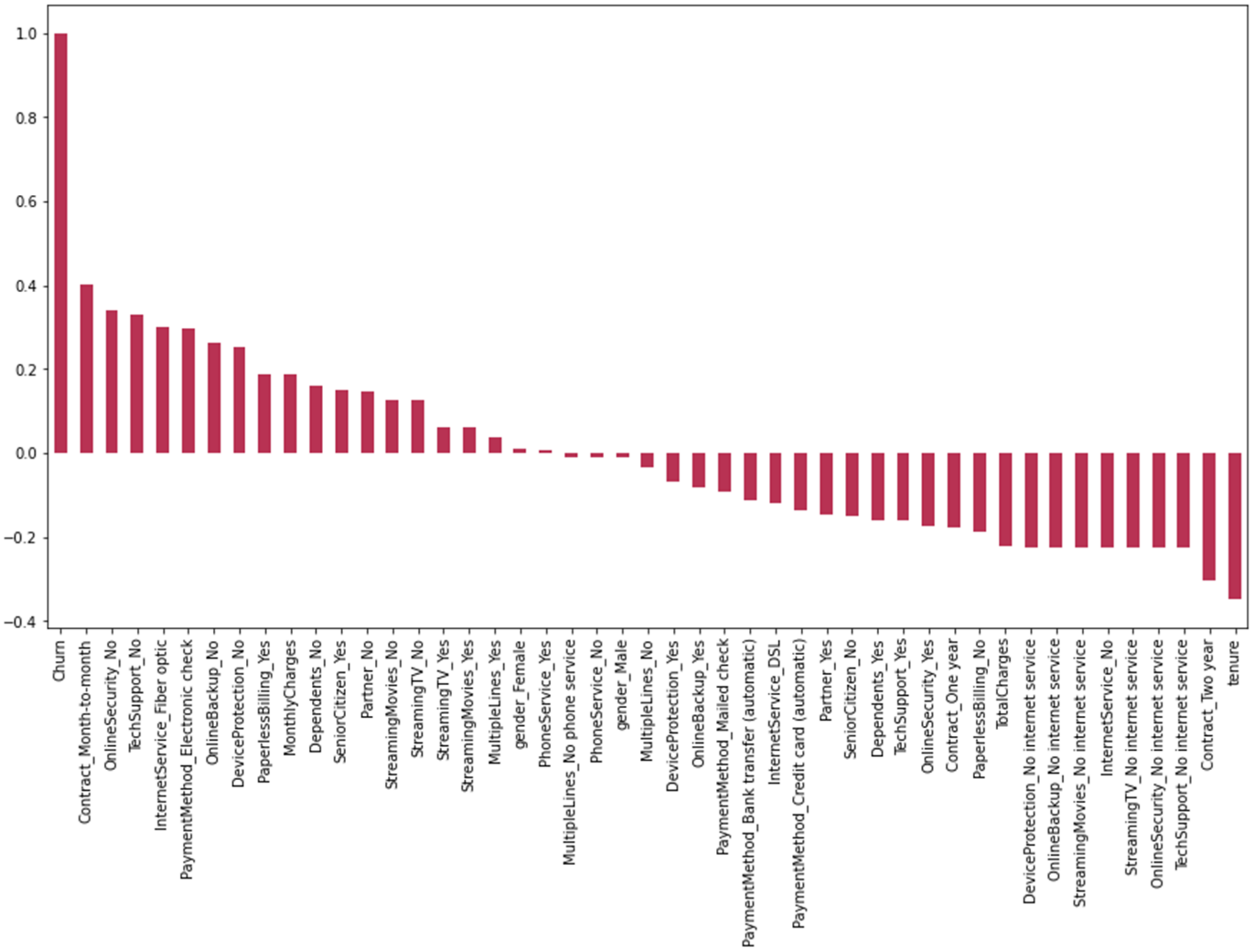


Figure 39: Churn correlation to all values

- Month to month contracts, absence of online security and tech support seem to be the most positively correlated values with churn ever, while tenure, two year contracts seem to be the most negatively correlated with churn. This means that likeliness of churn decreases with tenure, having a two year contract…and decreases in the case of having a month to month type of contracts…
- Having no internet, or related services in general, is negatively correlated to costumers' churn.
- Having an internet service or not, gender and having multiple lines seem to not be so correlated to churn. The possibility that dropping these features would be beneficial to the model should be tested.
- Customers might be unsatisfied with the quality of the fiber optic internet service, the company needs to address this issue.

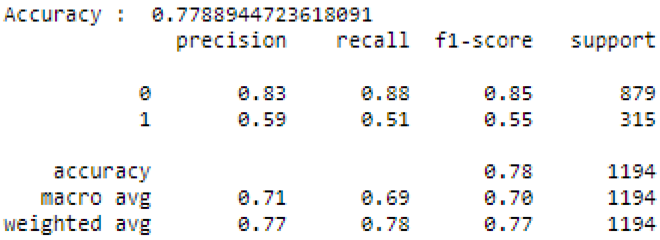
Now that we have formed a clear idea of the dataset, the next step would be to prepare it for modeling. That includes:

- Splitting the dataset to the target and label values
- Splitting it to training and testing datasets before any other form of preprocessing to avoid data leakage.
- One hot encoding the categorical attributes and using a min max scaler on numeric attributes.

Now for modelling, I fit the prepared dataset to 3 classifiers to evaluate their performance and choose the good – say right model.

i. RandomForestClassifier

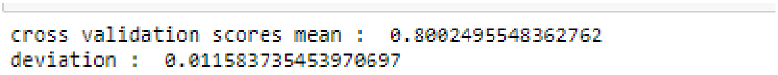
The classifier had one hyperparameter; random_state set to 42. The accuracy report:



```
Accuracy :  0.7788944723618091
              precision    recall  f1-score   support

           0       0.83      0.88      0.85       879
           1       0.59      0.51      0.55       315

    accuracy                           0.78      1194
   macro avg       0.71      0.69      0.70      1194
weighted avg       0.77      0.78      0.77      1194
```

Figure 40: RandomForestClassifier accuracy report

10 Stratified cross validation after Hyperparameters tuning:



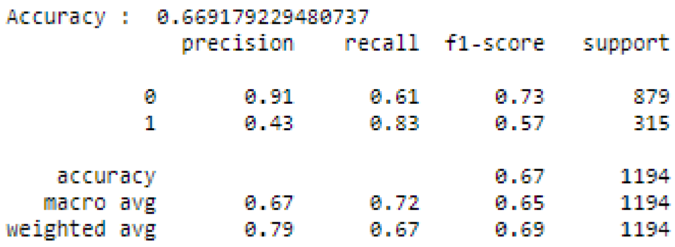
```
cross validation scores mean :  0.8002495548362762
deviation :  0.011583735453970697
```

Figure 41: RandomForestClassifier cross validation score after hyperparameter tuning

The deviation is extremely low – 1.15%, which means which means that this model has a very low variance; the model will perform more or less similar on all test sets.

ii. GaussianNB

The accuracy report:



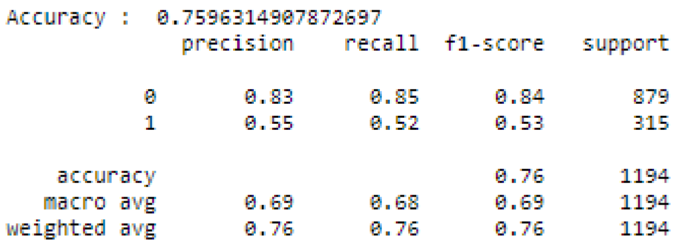
```
Accuracy :  0.669179229480737
              precision    recall  f1-score   support

           0       0.91      0.61      0.73       879
           1       0.43      0.83      0.57       315

    accuracy                           0.67      1194
   macro avg       0.67      0.72      0.65      1194
weighted avg       0.79      0.67      0.69      1194
```

Figure 42: GaussianNB accuracy report

iii. MLPClassifier

The accuracy report:

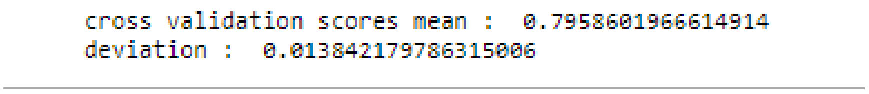
The classifier had one hyperparameter; the hidden layer size was set to (81, 1).



```
Accuracy :  0.7596314907872697
              precision    recall  f1-score   support

           0       0.83      0.85      0.84       879
           1       0.55      0.52      0.53       315

    accuracy                           0.76      1194
   macro avg       0.69      0.68      0.69      1194
weighted avg       0.76      0.76      0.76      1194
```

Figure 43: MLPClassifier accuracy report

10 Stratified cross validation after hyperparameters tuning:



```
cross validation scores mean :  0.7958601966614914
deviation :  0.013842179786315006
```

Figure 44: MLPClassifier cross scores after hyperparameter tuning

- Similarly to the RandomForestClassifier, the deviation is extremely low – 1.38%, this model as well will perform more or less similar on all test sets.

- RandomForestClassifier had seemingly better accuracy, precision and recall score, while MLPClassifier was similar to it, the GaussianNB performed quite poorly in comparison.

  The three models showed better accuracy with dropping the ‘PhoneService’ column than with keeping it, also they showed less accuracy while training them in with the ‘gender’ and ‘MultipleLines’ column, showing that the ‘PhoneService’ is better dropped.

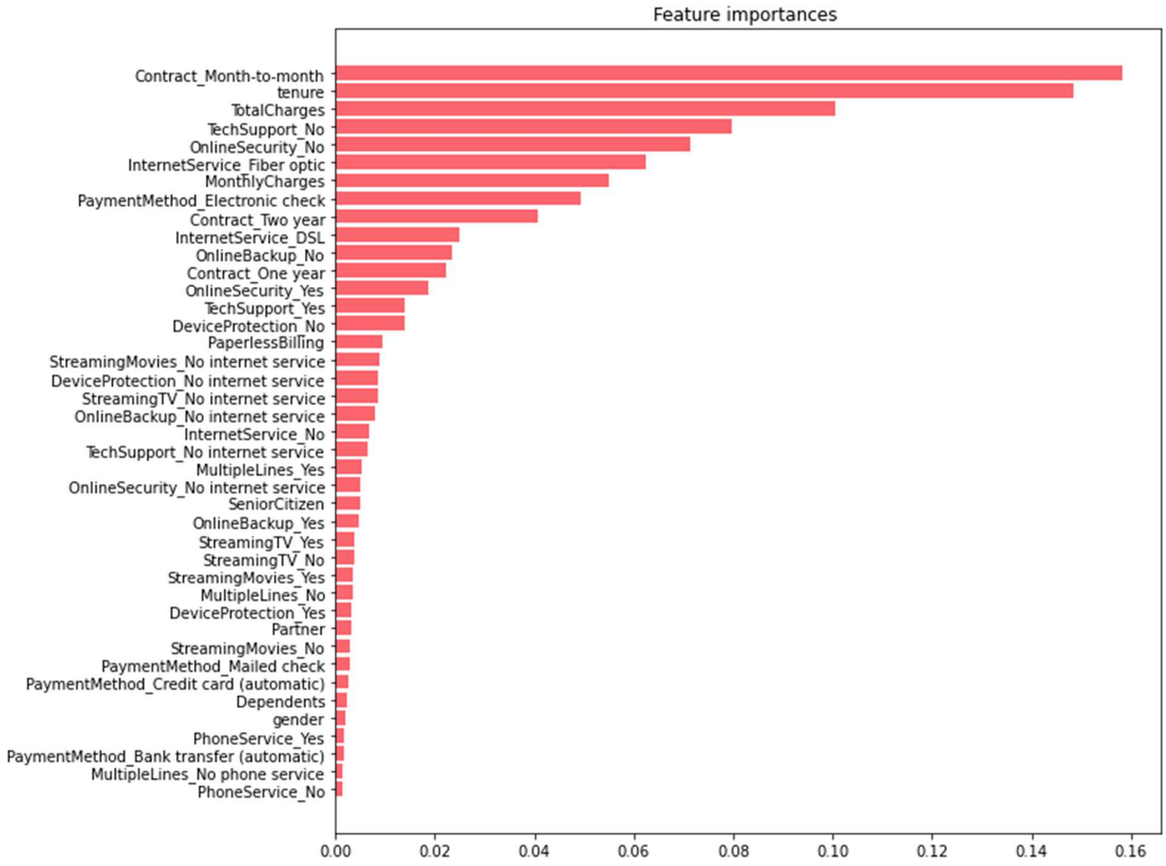


Figure 45: RandomForestClassifier feature by their importance

The plot shows that the ‘PhoneService’ values’ columns are the least important for the RandomForestClassifier decision.

Learning curves need to be analyzed next to choose the right model for our dataset.

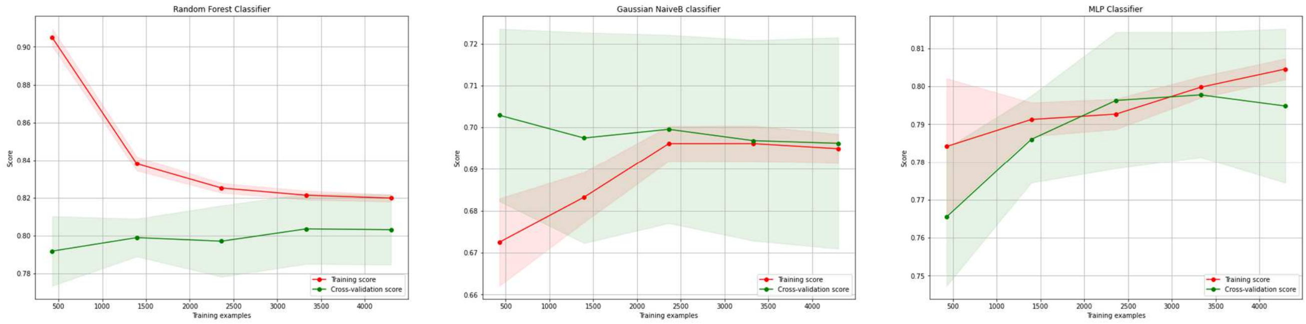


Figure 46: Classifiers' learning curves

- The RandomForestClassifier learning curve converges, which means it is a good fit, it is the most accurate model, it also indicates that the addition of more training examples doesn't improve the model performance on unseen data.
- It is clear the GaussianNB classifier is not the best choice to the dataset
- The MLPClassifier learning curve indicates that the classifier needs more data, as the accuracy is going up with the size of training.

The RandomForestClassifier seems to be the best choice.

II. The application set up

1. Used tools:

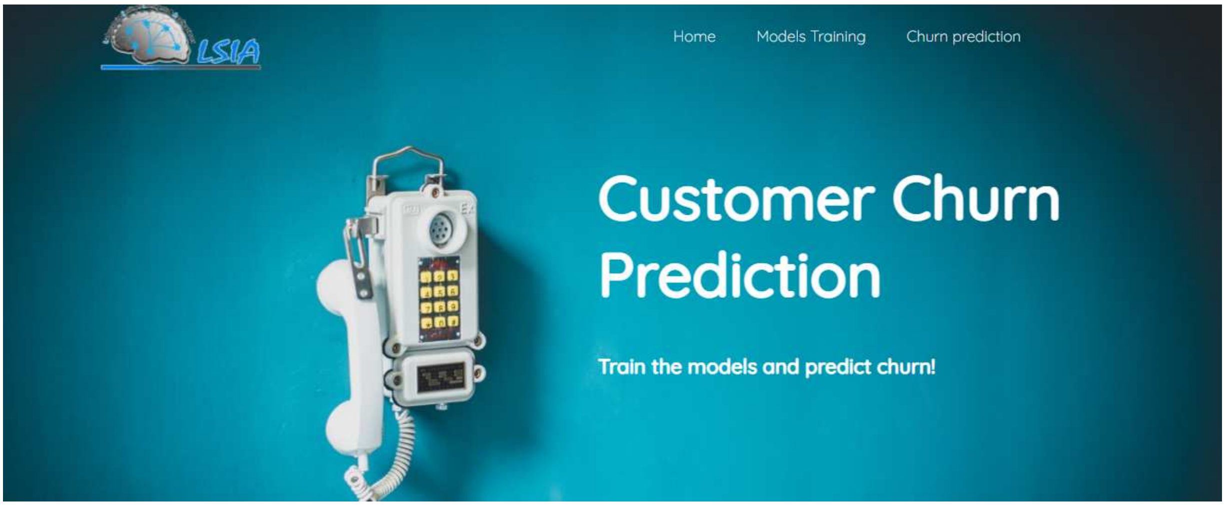
Now that the models are fitted, the next step is to deploy them into the application where they can be used to predict customers' decisions. The application also provides the possibility to visualize the models performance by displaying a confusion matrix of each model. The used tools to develop it:

- Pickle: Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it "serializes" the object first before writing it to file. Pickling is a way to convert a python object: list, dict…
- Flask: is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.
- HTML: Hypertext Markup Language is the most basic building block of the Web. It defines the meaning and structure of web content. Other technologies besides HTML are generally used to describe a web page's appearance/presentation (CSS) or functionality/behavior (JavaScript)
- CSS: Cascading Style Sheets is a style sheet language used for describing the presentation of a document written in a markup language such as HTML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript.
- JavaScript: programming language that conforms to the ECMAScript specification. JavaScript is high-level, often just-in-time compiled, and multi-paradigm. It has curly-bracket syntax, dynamic typing, prototype-based object-orientation, and first-class functions. JQuery is a JavaScript library that provides an Ajax framework and other utilities, with Ajax, web applications can send and retrieve data from a server asynchronously without interfering with the display and behavior of the existing page.

2. The application presentation:

This section is dedicated to the application’s interfaces presentation.

First the Homepage:

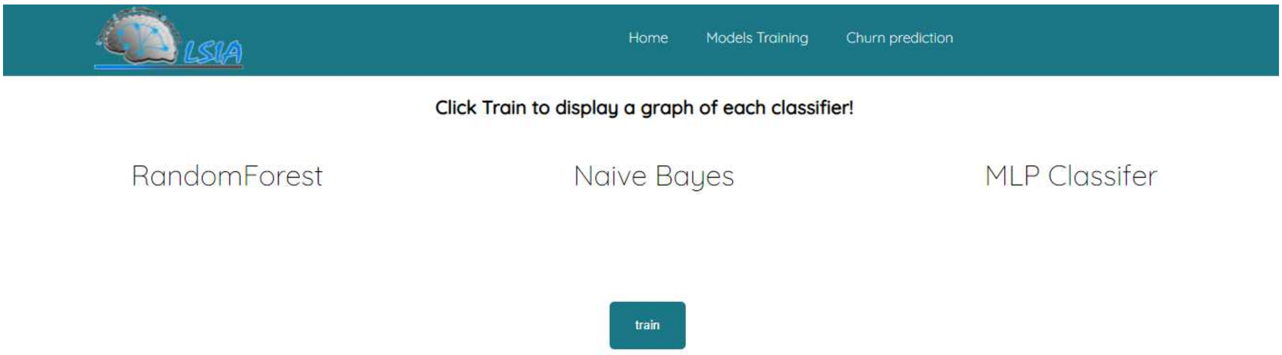


Customer Churn Prediction

Train the models and predict churn!

Figure 47: ChurnAPP homepage

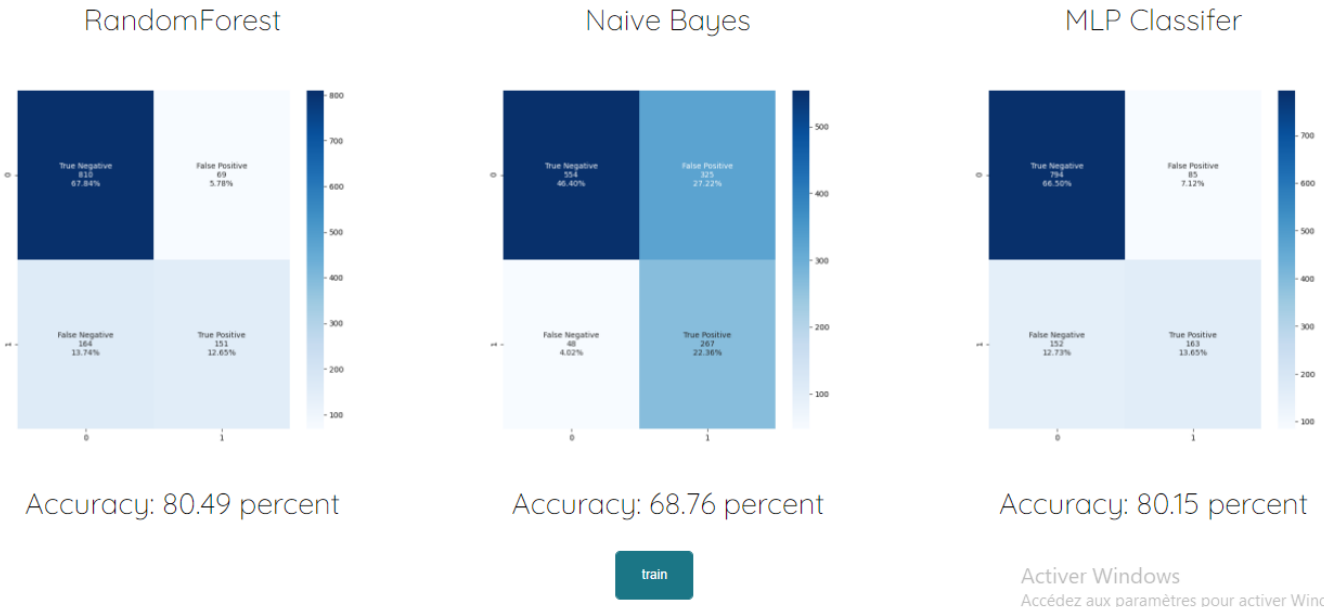
The second interface is for the models training:



Click Train to display a graph of each classifier!

Figure 48: Models training interface

After Clicking train, three graphs are displayed in the same interface:



Accuracy: 80.49 percent

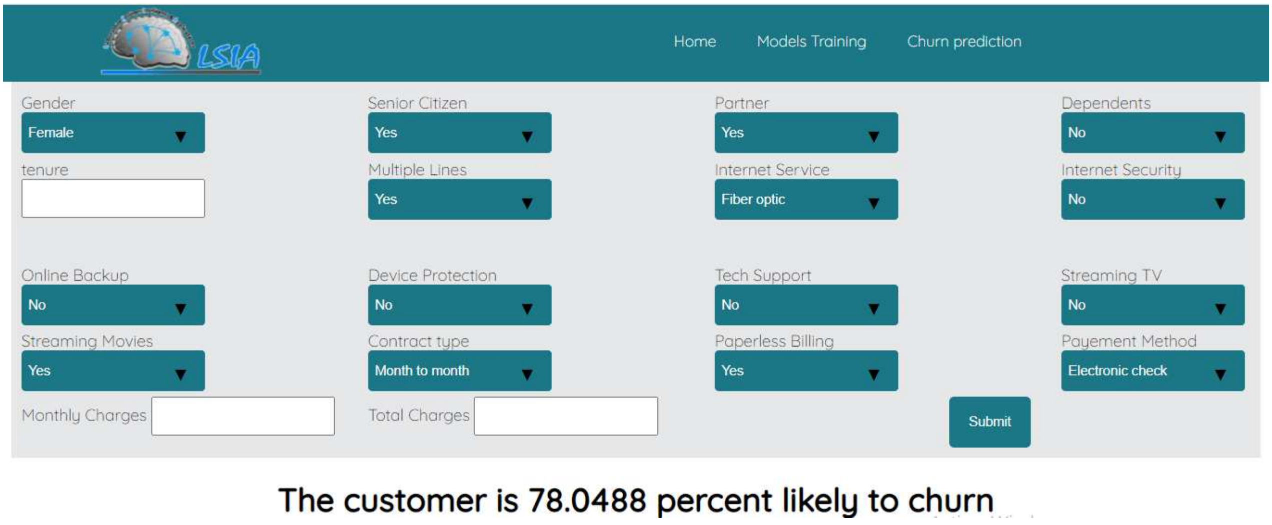
Accuracy: 68.76 percent

Accuracy: 80.15 percent

Figure 49: Displayed graphs

The displayed graphs show that the RandomForestClassifier is the right choice for this dataset since it performed on it better than the other two, since its f1-score is higher.

Lastly, the third interface provides the possibility to enter customer’ attributes in order to predict with a certain accuracy their decision:



Gender: Female
Senior Citizen: Yes
Partner: Yes
Dependents: No
tenure:
Multiple Lines: Yes
Internet Service: Fiber optic
Internet Security: No
Online Backup: No
Device Protection: No
Tech Support: No
Streaming TV: No
Streaming Movies: Yes
Contract type: Month to month
Paperless Billing: Yes
Payement Method: Electronic check
Monthly Charges:
Total Charges:

The customer is 78.0488 percent likely to churn

Figure 50: Churn prediction interface

Conclusion

This end-of-study project is an Exploratory Data Analysis of a telecommunication company's customers, well data, to understand it first then prepare it to modeling with machine learning algorithms and deploying the suitable ones in a web application that predicts these customers' decision.

It was a humble attempt in training models to predict future which has a million way to improve such us working more on hyperparameters tuning to improve the models performance and use some Data Augmentation techniques to generate new data and increase its diversity which might be beneficial to the models, especially the MLPClassifier as more data can result in a more skillful neural network model, as well as using more somewhat advanced techniques to explore and analyze the data.

This whole experience have been an opportunity to work on something, that I am not quite familiar with yet very enthusiastic about, that helped me to develop a great deal of analytical and researching skills over the course of it and explore relatively new concepts and learn how to put them into application.

Bibliography

Data Preparation for Machine Learning (2020), by Jason Brownlee

Les techniques d'apprentissage automatique at d'aide à la décision, by PR. Aicha MAJDA

Principles of Data Wrangling (2017), by Joseph M. Hellerstein, Tye Rattenbury, Jeffrey Heer, Sean Kandel, Connor Carreras.

Applied Predictive Modeling (2013), by max Kuhn kjell Johnson

Data Science Handbook (2017), by Jake VanderPlas

Data visualization a practical introduction (2018), by Kieran Healy

Webography

- https://docs.python.org/3.9/tutorial/
- https://flask.palletsprojects.com/
- https://scikit-learn.org/
- https://matplotlib.org/
- https://seaborn.pydata.org/
- https://numpy.org/
- https://pandas.pydata.org/
- https://www.kaggle.com/

References

[1] C. Petrov, "Impressive Big Data Statistics for 2021," *techjury,* June 2021.

[2] "cloudmoyo," 2020. [Online].

[3] "The world's most valuable resource is no longer oil, but data," *The Economist,* 2017.

[4] J. Milligan, "Data is Only as Valuable as the Insights it Informs," 2019.

[5] "Data Analytics Market to Hit USD 132.90 Billion by 2026," Market Research Future, 08 February 2021. [Online].

[6] "Structured vs. Unstructured Data: A Complete Guide," talend. [Online].

[7] Z. Plaksij, "CUSTOMER CHURN: 12 WAYS TO STOP CHURN IMMEDIATELY," superoffice, 04 May 2021. [Online].

[8] U. A. Suresh Kumar Mukhiya, Hands-On Exploratory Data Analysis with Python, Packt, 2020.

[9] C. K. Rachel Hornay, "Data Analytics vs Data Analysis: What's The Difference?," bmc blogs, 08 January 2021. [Online].

[10] S. K. Arora, "What is Data Analysis? Methods, Techniques & Tools," hackr.io, 20 April 2021. [Online].

[11] "e-Handbook of Statistical Methods," NIST/SEMATECH, 30 October 2013. [Online].

[12] "The Best Data Science Methods For Predictive Analytics," conceptatech, 29 November 2017. [Online].

[13] J. Brownlee, Data Preparation for Machine Learning, Machine Learning Mastery, 2020.

[14] E. Escott, "A guide to narrow, general, and super artificial intelligence," codebots, 24 October 2017. [Online].

[15] Hwawei ICT, 2020.

[16] A. Navlani, "Naive Bayes Classification using Scikit-learn," datacamp, 4 December 2018. [Online].

[17] "All about Naive Bayes," toward data science, 08 October 2018. [Online].

[18] P. Vadapalli, "Naive Bayes Classifier: Pros & Cons," upgrad, 11 December 2020. [Online].

[19] "Classification Algorithms - Random Forest," tutorialspoint, 2018. [Online].

[20] "Random Forest Algorithm- An Overview," mygreatlearning, 19 February 2020. [Online].

[21] "Wikipedia," 12 June 2021. [Online].

[22] "The differences between Artificial and Biological Neural Networks," 2018. [Online].

[23] "Multilayer Perceptron (MLP)," data science bootcamp, 22 December 2018. [Online].

[24] G. Ciaburro and B. Venkateswaran, Neural Networks with R, Packt, 2017.

[25] "Evaluating a machine learning model," jeremyjordan, 21 July 2017. [Online].

[26] A. Bora, "Using Learning Curves – ML," geeksforgeeks, 07 July 2020. [Online].

[27] R. Ng, "ML - Learning Curve," RichieNg, 2017. [Online].

[28] B. Ghojogh and M. Crowley, The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial., 2019.

[29] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Machine Learning Mastery, 23 May 2018. [Online].

[30] "Evaluating Machine Learning Models using Hyperparameter Tuning," analytics vidhya, 12 April 2021. [Online].