U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Supervised Learning Telco Customer Churn

IART Project 2

Carlos Gomes – up201906622

Domingos Santos – up201906680

Filipe Pinto – up201907747

Specification

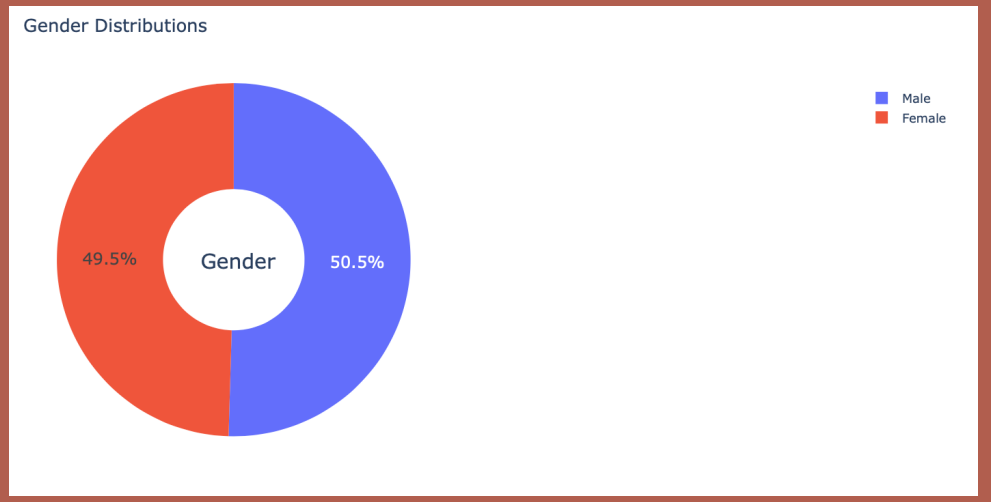
- It is important for a company to retain customers in order to maintain or even increase profit, so it might be very useful to predict their behaviour. To do that we need to make a market research to answer some questions.
- So, given a dataset with information about telco customers we want to predict if a customer will churn or not, according to the percentage of churn in the dataset and if that number is affected by any other variable such as gender, services subscribed or even the charges of the customer.
- Other important analysis for the company are the profit evaluation such as the most profitable service or feature and the ones not sot profitable.
- All of this questions/doubts should be after the study of the dataset and that'ś the main goal of this project.

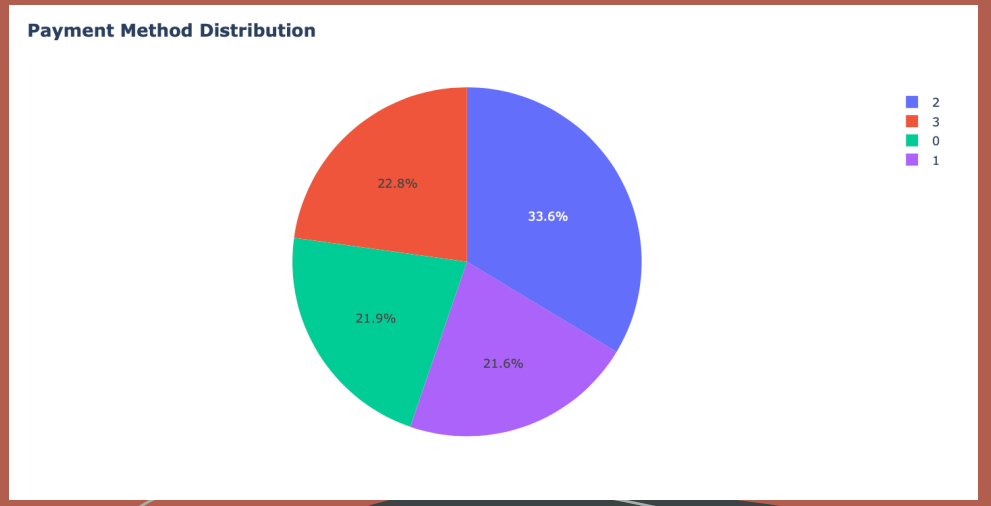
Related Work

- https://moodle.up.pt/ (course files)
- https://www.kaggle.com/datasets/easonlai/sample-telco-customer-churn-dataset
- https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Dataset Analysis

- After analysing our data, we concluded that gender distribution is equal for both genders, however, payment method distribution is not uniform, verifying distinct distribution;





Data Pre-processing

- **Load data;**
- **Missing and Repeated values**
  - There are not any missing values in our dataset. If we needed to eliminate any duplicated value, it only could be a repeated customerid row, due to the types and values of each column (most of them can be represented as booleans so there must exist duplicate values in those columns) and, as we can see, the number of unique id's match exactly the number of rows so there is no repeated value;
- **Drop Columns**
  - We can drop column customerID because it has no influence in churn value;
- **Removing NA answers**
  - Many columns ahve and NA value meaning that the customer did not answer if they subscribe a type of service or if they use specific service for a given purpose. We assume that if the answer is null that answer is a No. For example, when a customer does not say of he/she subscribes to an additional online security service provided by the company we assume that the answer if a No;

Data processing

- There are two datasets available let's call the main(with 7011 entries) and the other test (21 entries). For our test/train model we will use the main one. Before any attempt of method implementation we need to do some data analysis and pre-processing. We first need to load the data. Then we may print some information about the dataset to get used to the way the information is organized and to know what our next step should be, such as statistics and possible missing and repeated values. The following topics indicate the description of each column in the dataset.

Data processing

Column Description:

- customerID: A unique ID that identifies each customer.
- gender: The customer's gender: Male (1), Female (0).
- SeniorCitizen: Indicates if the customer is 65 or older: No (0), Yes (1).
- Partner: Service contract is resold by the partner: No (0), Yes (1).
- Dependents: Indicates if the customer lives with any dependents: No (0), Yes (1).
- Tenure: Indicates the total amount of months that the customer has been with the company.
- PhoneService: Indicates if the customer subscribes to home phone service with the company: No (0), Yes (1).
- MultipleLines: Indicates if the customer subscribes to multiple telephone lines with the company: No (0), Yes (1).
- InternetService: Indicates if the customer subscribes to Internet service with the company: No (0), DSL (1), Fiber optic (2).
- OnlineSecurity: Indicates if the customer subscribes to an additional online security service provided by the company: No (0), Yes (1), NA (2).
- OnlineBackup: Indicates if the customer subscribes to an additional online backup service provided by the company: No (0), Yes (1), NA (2).
- DeviceProtection: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: No (0), Yes (1), NA (2).
- TechSupport: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: No (0), Yes (1), NA (2).
- StreamingTV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: No (0), Yes (1), NA (2). The company does not charge an additional fee for this service.
- StreamingMovies: Indicates if the customer uses their Internet service to stream movies from a third party provider: No (0), Yes (1), NA (2). The company does not charge an additional fee for this service.
- Contract: Indicates the customer's current contract type: Month-to-Month (0), One Year (1), Two Year (2).
- PaperlessBilling: Indicates if the customer has chosen paperless billing: No (0), Yes (1).
- PaymentMethod: Indicates how the customer pays their bill: Bank transfer - automatic (0), Credit card - automatic (1), Electronic cheque (2), Mailed cheque (3).
- MonthlyCharges: Indicates the customer's current total monthly charge for all their services from the company.
- TotalCharges: Indicates the customer's total charges.
- Churn: Indicates if the customer churn or not: No (0), Yes (1).

Data processing

- Train and split our data;
- To contribute for an equally to model fitting, we used automatic scales;

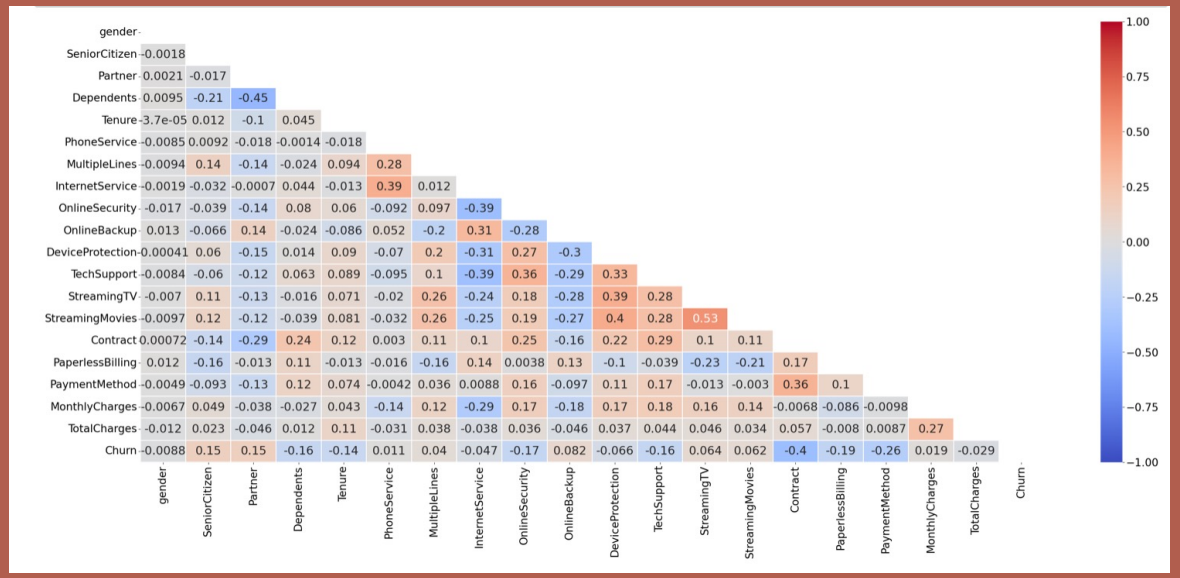
Tools and algorithms

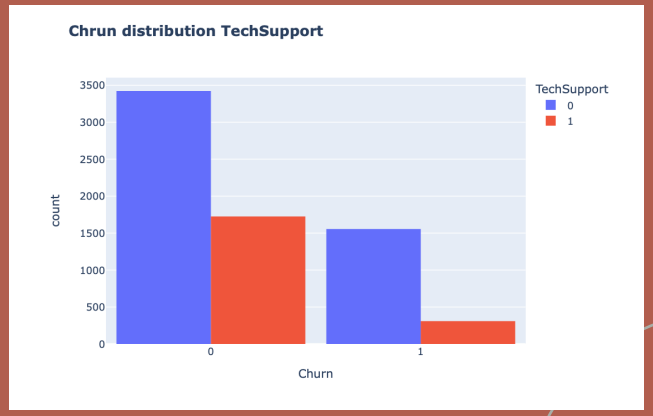
- For this assignment we will use some python tools and libraries also used in classes which are:
- numpy (library used for working with arrays);
- pandas (data science/data analysis and machine learning tasks);
- cikit-learn (machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction);
- matplotlib (comprehensive library for creating static, animated, and interactive visualizations in Python);
- seaborn (uses Matplotlib underneath to plot graphs),
- plotpy;
- Our work will be developed in a python notebook- In our case Jupyter Notebook, so this packages come standard with the Anaconda python distribution;
- To reach the main goal, we need to implement some classification algorithms for supervised learning such as Support Vector Machine, K-Nearest Neighbours or Decision Tree Classification;

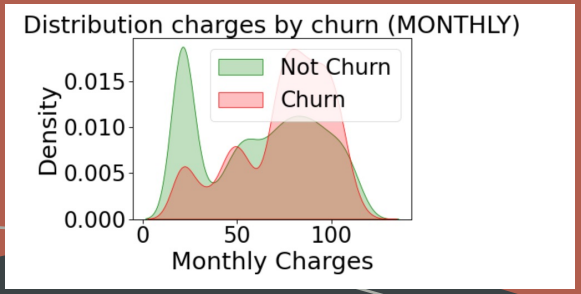
Data Visualisation

- We analysed some of the data provided and made a few changes in order to easily have a better aproach for our solution of the problem. Now we will provide some different types of graphics and plots for a user friendly comprehension of the data to study.

Data Visualisation – Some Examples





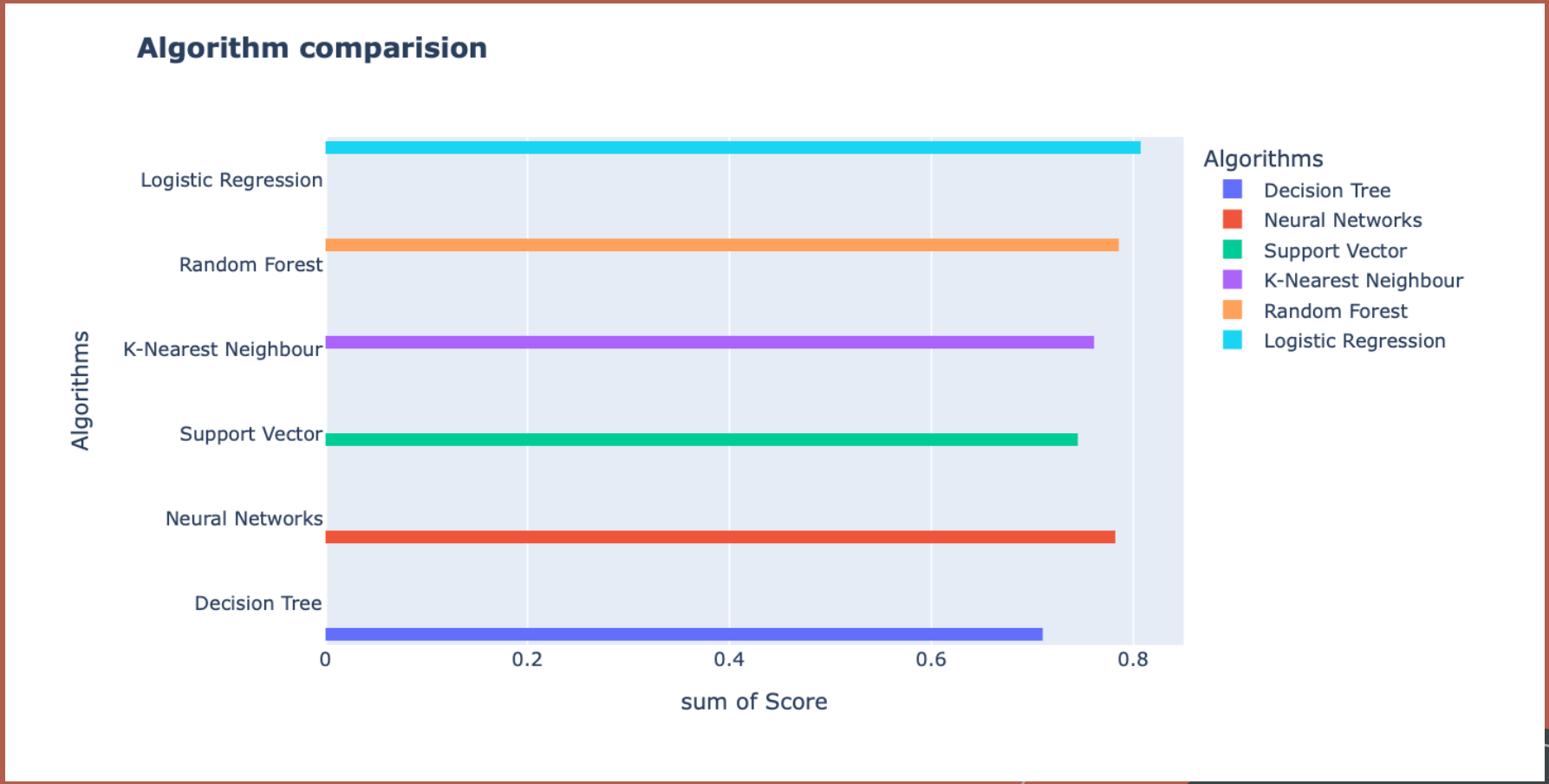


Developed Algorithms - Supervised Learning

- Decision Tree Classifier: Using a DTC, we want to predict the value of a target variable by learning simple decision rules from our data features;
- Neural Networks: By processing exemples, our algorithm learns what it should do, forming probability-weighted associations between “input” and “result”, stored in the data structure of the net itself;
- Support Vector Classification: Learning problem is formulated as a convex optimization problem and is robust to noise;
- K-nearest Neighbors: assuming the similarity between data and available cases, put a new case into a certain category that is most similar to the available categories;
- Logistic Regression: Binary method used to predict a binary outcome, in this case, yes or no, based on prior observations of a data set;
- Random Forest: is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.;

NOTE: To improve our results, we also used Parameter Tuning using GridSearchCV;

Developed Algorithms – Results Comparison



Conclusion

- All in all, with this project we learned a little bit more about Machine Learning, specially about Supervised Learning.
- For this we were first introduced to data pre-processing and visualization, which means becoming familiar with the dataset proposed for study and all the data manipulation needed for our results have a reliable meaning. Then, we started to perform some models using the algorithms provided by the libraries used, which was something that we learn in this project, the hability to work with some python libraries and tools which were unknown for us some weeks ago.
- Using the obtained score for each algorithm we also plotted their performance. Even though, we have achieved good results, these could've been better if more data have been manipulated, for example removing some columns which, according to the data visualization analysis, were not important to affect churn.