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

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;

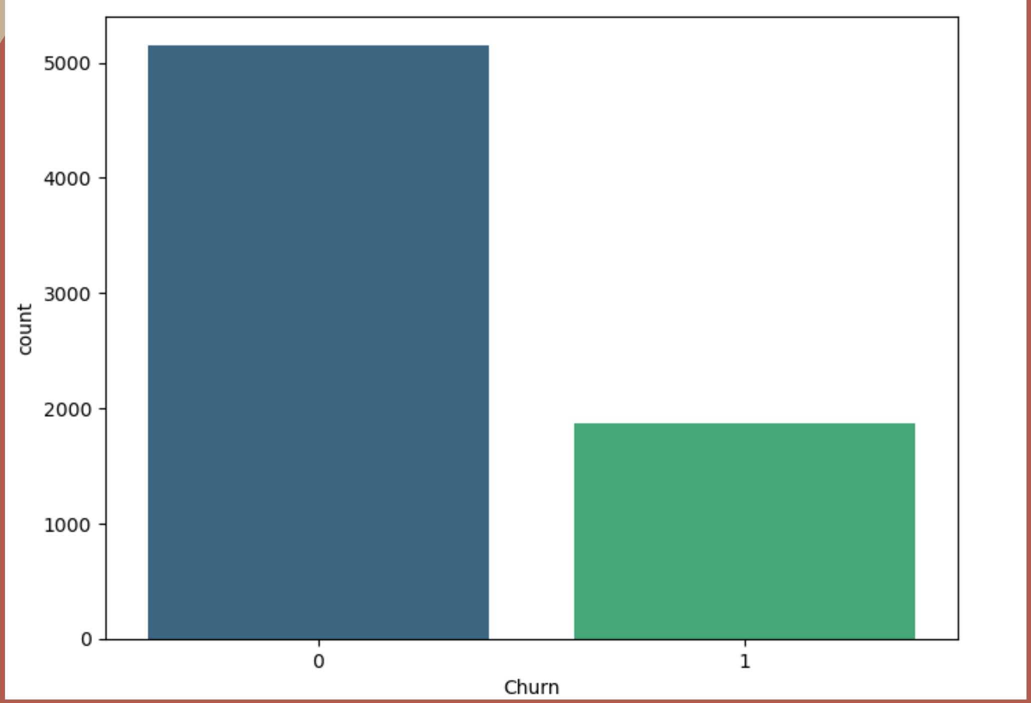
Tools and algorithms

- 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;

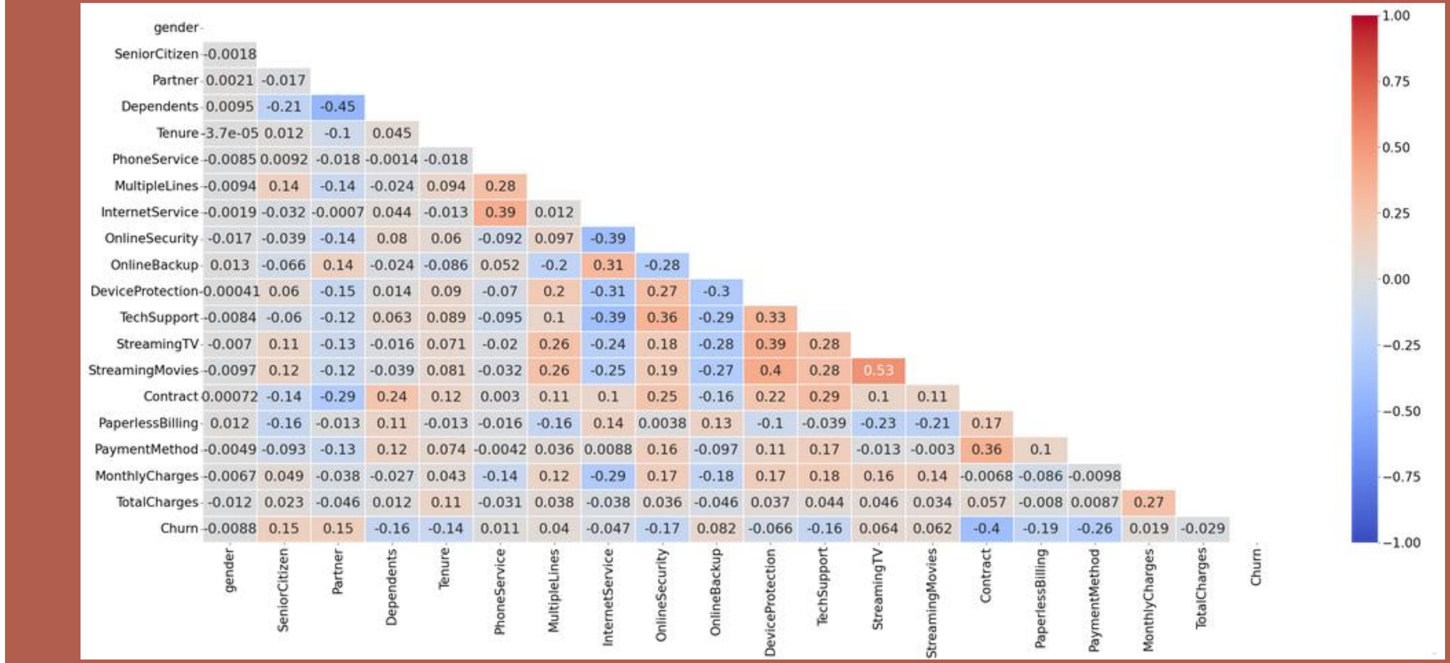
Data Visualization

- We analysed some of the data provided and made a few changes, in order to easily have a better approach for our solution of the problem. Then we created some graphics and plots, of different types and with different data, for a user friendly comprehension of the data to study.

Data Visualization – Some Examples



Churn distribution



	gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
gender																				
SeniorCitizen	-0.0018																			
Partner	0.0021	-0.017																		
Dependents	0.0095	-0.21	-0.45																	
Tenure	-3.7e-05	0.012	-0.1	0.045																
PhoneService	-0.0085	0.0092	-0.018	-0.0014	-0.018															
MultipleLines	-0.0094	0.14	-0.14	-0.024	0.094	0.28														
InternetService	-0.0019	-0.032	-0.0007	0.044	-0.013	0.39	0.012													
OnlineSecurity	-0.017	-0.039	-0.14	0.08	0.06	-0.092	0.097	-0.39												
OnlineBackup	0.013	-0.066	0.14	-0.024	-0.086	0.052	-0.2	0.31	-0.28											
DeviceProtection	-0.00041	0.06	-0.15	0.014	0.09	-0.07	0.2	-0.31	0.27	-0.3										
TechSupport	-0.0084	-0.06	-0.12	0.063	0.089	-0.095	0.1	-0.39	0.36	-0.29	0.33									
StreamingTV	-0.007	0.11	-0.13	-0.016	0.071	-0.02	0.26	-0.24	0.18	-0.28	0.39	0.28								
StreamingMovies	-0.0097	0.12	-0.12	-0.039	0.081	-0.032	0.26	-0.25	0.19	-0.27	0.4	0.28	0.53							
Contract	0.00072	-0.14	-0.29	0.24	0.12	0.003	0.11	0.1	0.25	-0.16	0.22	0.29	0.1	0.11						
PaperlessBilling	0.012	-0.16	-0.013	0.11	-0.013	-0.016	-0.16	0.14	0.0038	0.13	-0.1	-0.039	-0.23	-0.21	0.17					
PaymentMethod	-0.0049	-0.093	-0.13	0.12	0.074	-0.0042	0.036	0.0088	0.16	-0.097	0.11	0.17	-0.013	-0.003	0.36	0.1				
MonthlyCharges	-0.0067	0.049	-0.038	-0.027	0.043	-0.14	0.12	-0.29	0.17	-0.18	0.17	0.18	0.16	0.14	-0.0068	-0.086	-0.0098			
TotalCharges	-0.012	0.023	-0.046	0.012	0.11	-0.031	0.038	-0.038	0.036	-0.046	0.037	0.044	0.046	0.034	0.057	-0.008	0.0087	0.27		
Churn	-0.0088	0.15	0.15	-0.16	-0.14	0.011	0.04	-0.047	-0.17	0.082	-0.066	-0.16	0.064	0.062	-0.4	-0.19	-0.26	0.019	-0.029	

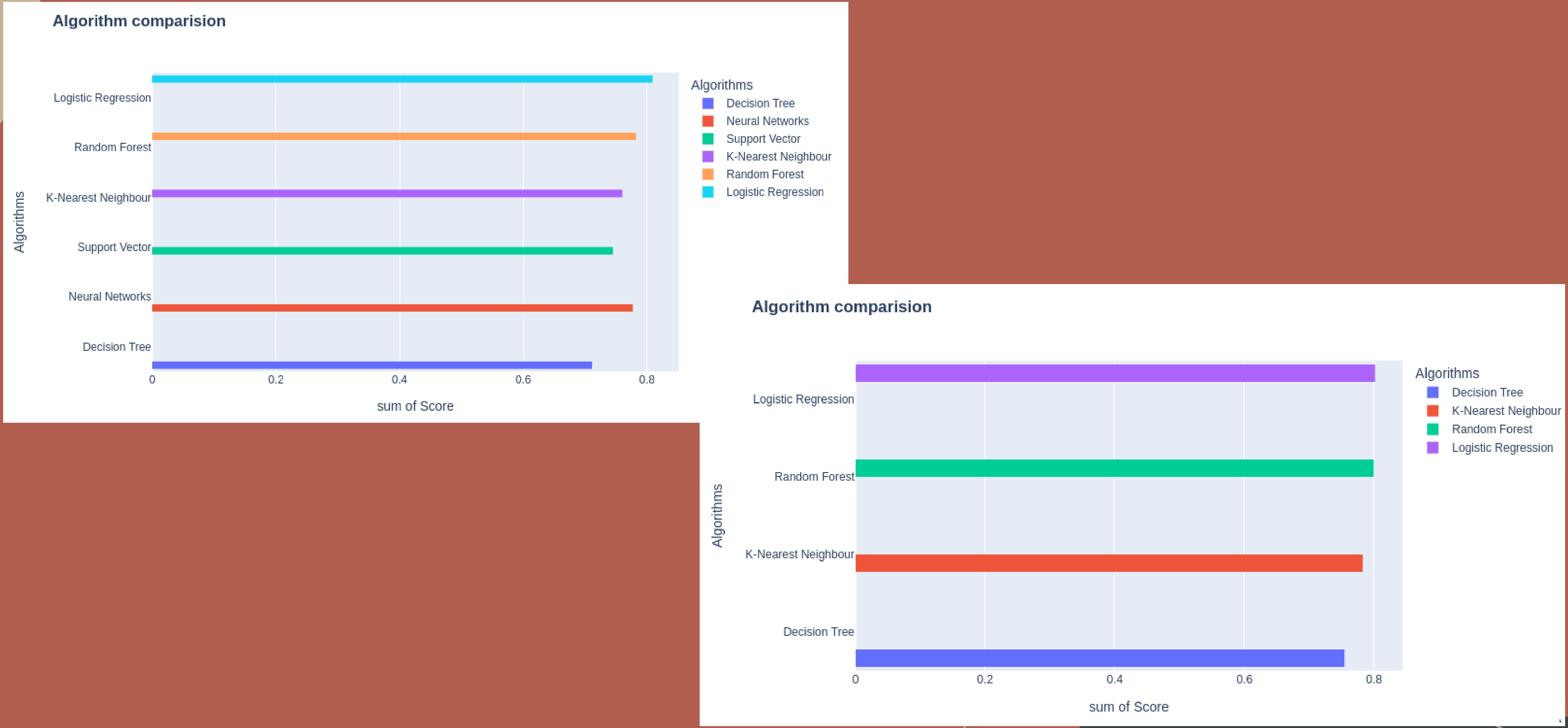
Correlation HeatMap

Many other examples can be found in the notebook

Developed Algorithms - Supervised Learning

- For our datset training we used 6 models: Decision Tree Classifier, Neural Networks, Support Vector Classification, K-Nearest Neighbors, Logistic Regression and Random Forest.
- In a first approach we used the default values for the parameters used in these algorithms (results and comparision in the next slide).
- Then we used GridSearch to find which parameter combination best suited in our model so that the accuracy score could be higher.
- In addiction, to each of the previous approaches we implemented a Voting Classifier. To do that we picked the 3 algorithms with the highest score(for each situation) and then applied the Voting Classifier which trains an ensemble of numerous models, in our case those top-3.

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.