**Customer Churn Machine Learning Project**

* **Introduction:**

Customer churn identifies the percentage of customers that are no longer using a certain company's product or service or chose to switch to an alternative company rival to the one he used to deal with. This range of customers is evaluated through a certain period of time.

Handling this issue, especially in the telecommunication industry, is crucial to a company’s stability and growth due to the fact that clients are companies’ greatest assets and keeping them is more profitable than acquiring new ones. So even though acquisition can be viewed as detrimental for company growth, customer retention might be the key factor for long-term success of the organization.

Throughout this project, we are aspiring for the most suitable solution to customer churn by applying the latest machine learning technologies in order to get the most accurate result possible utilizing a near perfect model referring to customer analysis behavior. Regarding our project plan, we implemented the cross-industry process for data mining (CRISP-DM) method: a process model with six phases that naturally describes the data science life cycle.

* **Business understanding**

Presently, mobile telecommunication is the most relevant way of communication around the planet, that's why even the biggest telecommunication companies are facing a huge problem due to the big saturation of the market which made them trying hardly to keep returning customers instead of finding new ones. This snag made the rivalry between telecommunication companies more and more intense especially when the algorithms and the models of predicting customer behaviours are invented. This new investment in customer churn issue gave a chance to take a very close look into customer behaviour which is a necessity nowadays in the telecommunication market as it could end up affecting the revenue numbers and influence policy decisions.

* **Determine business objectives :**

The main objective is reaching a model with the highest efficiency on the market.

We went through several solutions as comparing and analysing other company’s costumer churn and extracting insights in order to reach better solutions and come up with a more effective model. Being able to learn about the customer’s behaviour faster than the competitor is our main and prior goal.

We will be analysing the common and obvious customer tendencies that would lead to the churn and predicting the kind of subscribers who are more likely to switch the operator or to definitely quit and this by using machine learning tools. We are aiming to develop a model that predicts the customers with the higher probability to switch lines.

By gaining information about customers from individual demographics to details of usage of service, we’d be able to recognize which services are less likely to be consumed which lead to implement effective retention strategies by identifying the services that would fulfill customer’s preferences. As a result, we’re looking forward to attempt a growth rate higher than churn rate and decreasing more than 2% churn which is equivalent to 10% reduction in costs.

* **The AssessSituation:**

We are a group of a data science students, are trying to do churn analysis in order to come out with solutions that can identify the customers that will eventually decline or switch to another company. Using the database named "teleco-customer-churn", we will start firstly by the data preparation in order to reach a better understanding of the data by applying multiples methods such as K-Nearest Neighbours, Tree classifier and lofitic regression referring to cross validation for best practices. Using Python as the program language, Anaconda Jupiter as a platform implementing the Pandas, sickit.learn, Matplot.lib, NumPy and SEABORN libraries.

This project will normally be achieved by the 6th of January, we will deliver a report of [ 20-30] pages and a notebook well commented and detailed. The data base we are working with contains at least 2129 rows, 33 columns and 1Mo for its size. In the next step, data understanding, we will be cleaning the database for an accurate model.

When it comes to the risks of this project, the data base understanding phase can be difficult to do perfectly at the first try giving the different types of columns we have to work with.

* **Determine data mining goals :**

When it comes to the data mining goals, we will be comparing models based on the F1 score to resolve the classification issues in order to reach the most accurate and effective method and model giving us the perfect solution for the customer churn problem. We are aiming for a F1 score in the interval [0. 97..1].



* Data understanding :

The *Telco customer churn* data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer and Churn information.

* Initial data collection

The dataset we are studying in this project is *Telco customer churn*. This dataset is published online with a public access for learning purposes.

* Data description

The format of the dataset we are putting in use is Comma Separated Values (CSV). It contains 7043 rows and 21 columns putting it at a size of 1.1 M. The column types vary from 4 numerical columns to 17 String values. The dataset can bedivided to 3 sections:

* Demographics
* Services
* Status

Demographics

CustomerID(String): A unique ID that identifies each customer.

Gender(String: Categorical): The customer’s gender: Male, Female

Senior Citizen(String: Categorical) : Indicates if the customer is 65 or older: Yes, No

Partner(String: Categorical):Indicates the customer has a partner: Yes, No.

Dependents(String: Categorical) : Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Services

Tenure in Months(int64): Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Phone Service(String: Categorical): Indicates if the customer subscribes to home phone service with the company: Yes, No

Multiple Lines(String: Categorical): Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No, No phone service

Internet Service(String: Categorical): Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic.

Online Security(String: Categorical): Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No, No Internet service.

Online Backup(String: Categorical): Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No, No Internet service.

Device Protection Plan(String: Categorical): Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No, No Internet service

Streaming TV(String: Categorical): Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No, No Internet service.

Streaming Movies(String: Categorical): Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No, No Internet service.

Contract(String: Categorical): Indicates the customer’s current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing(String: Categorical): Indicates if the customer has chosen paperless billing: Yes, No

Payment Method(String: Categorical): Indicates how the customer pays their bill: Bank transfer(Automatic),Credit Card (Automatic), Mailed Check, Electronic Check.

Monthly Charge(Float64): Indicates the customer’s current total monthly charge for all their services from the company.

Total Charges(String): Indicates the customer’s total charges, calculated to the end of the quarter specified above.

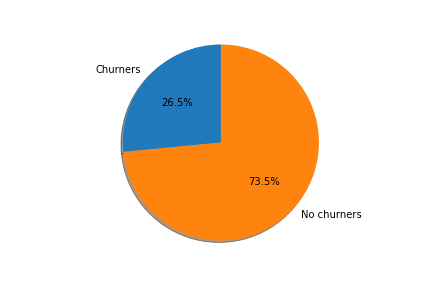
Status

Churn Label(String: Categorical): Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

* Data Exploration

In this part, we explored the dataset categorical features and these are the observations and plots we detected and created.

Churners : from 7043 clients, only 1869 are churners.



Gender :

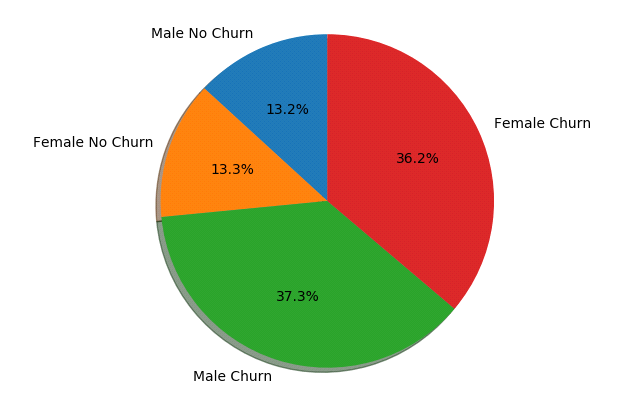
Male no churners : 2625

Female no churners : 2549

Male churners : 930

Femalechurners : 939

* Gender has no influence over the results.



Senior Citizen :

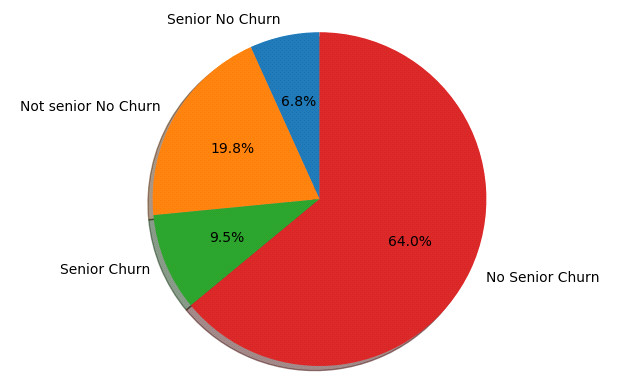
Yes no churners : 666

No no churners : 4508

Yes churners : 476

No churners : 1393

* The younger generation tends to churn more.



Partner :

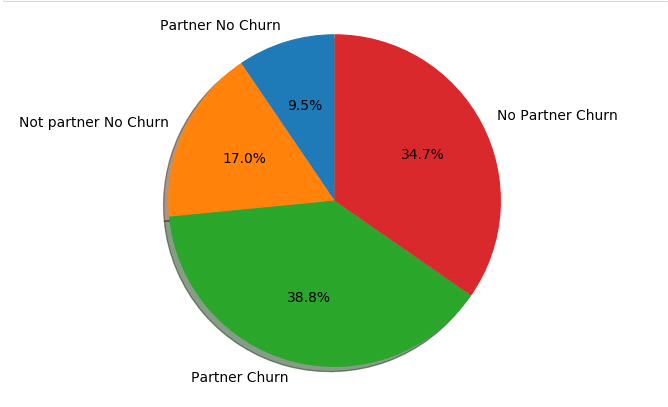
Yes no churners : 2733

No no churners : 2441

Yes churners : 669

No churners : 1200

* Single cutomers have the tendency to churn more.



Dependant :

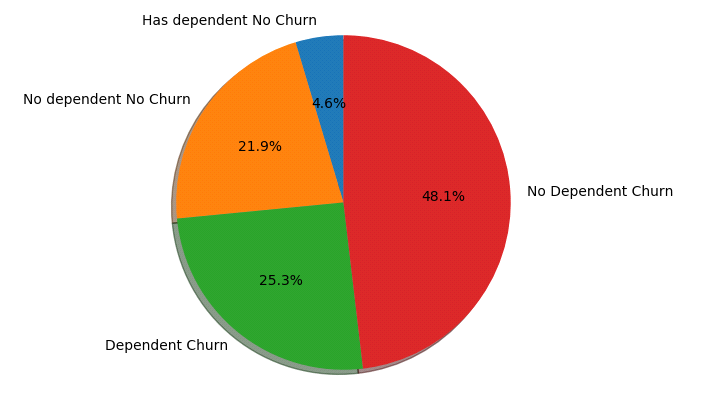
Yes no churners : 1521

No no churners : 3653

Yes churners : 106

No churners : 1763

* Customers without kids tend to churn more



Paperless billing :

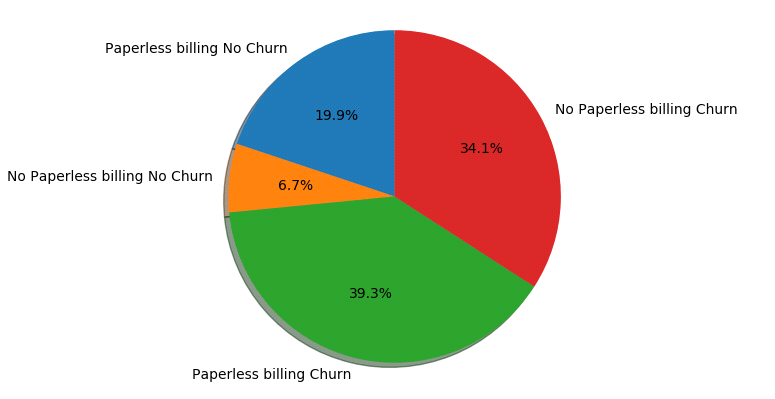
Yes no churners : 2771

No no churners : 2403

Yes churners : 1400

No churners : 469

* Customers with paperless billing churn more



Multiple Lines :

No no churners : 2541

Yes no churners : 2121

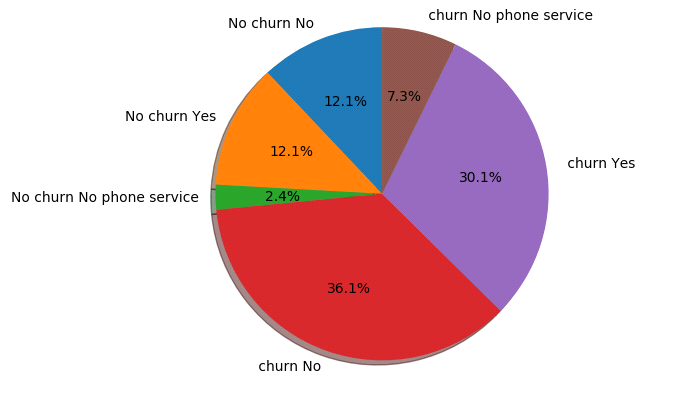
No phone service no churners : 512

No churners : 849

Yes churners : 850

No phone service churners : 170

* Multiple Lines has no influence over the results.



Internet Service :

DSL no churners : 1962

Fiber optic no churners : 1799

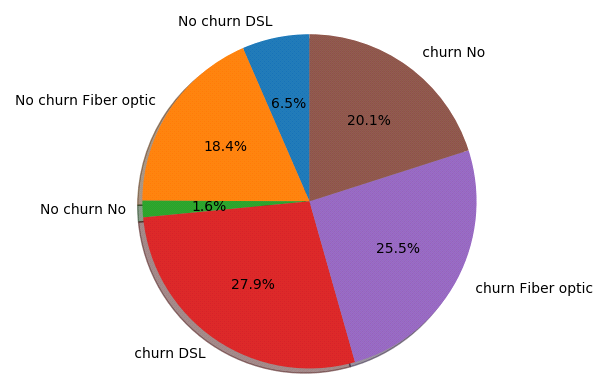
No no churners : 1413

DSL churners : 459

Fiber optic churners : 1297

No churners : 113

* Customers with fiber optic churn more



Online Security:

Yes no churners : 1724

No no churners : 2037

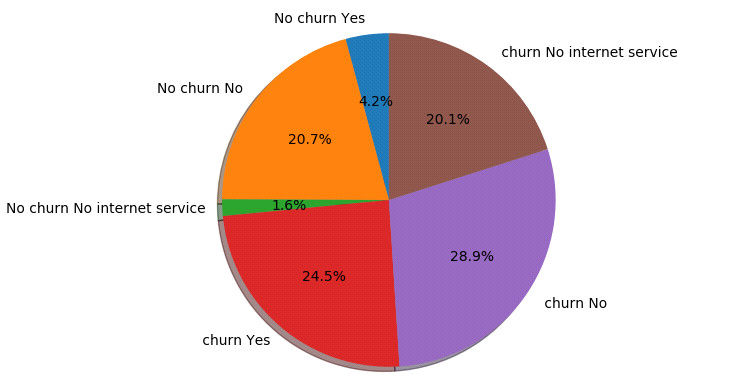
No internet service no churners : 1413

Yes churners : 295

No churners : 1461

No internet service churners : 113

* Customers without online security churn more



Online Backup :

Yes no churners : 1906

No no churners : 1855

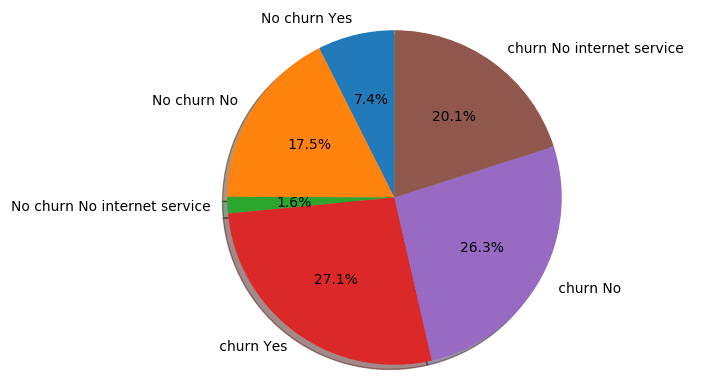
No internet service no churners : 1413

Yes churners : 523

No churners : 1233

No internet service churners : 113

* Customers without online backup churn more



Device Protection:

Yes no churners : 1877

No no churners : 1884

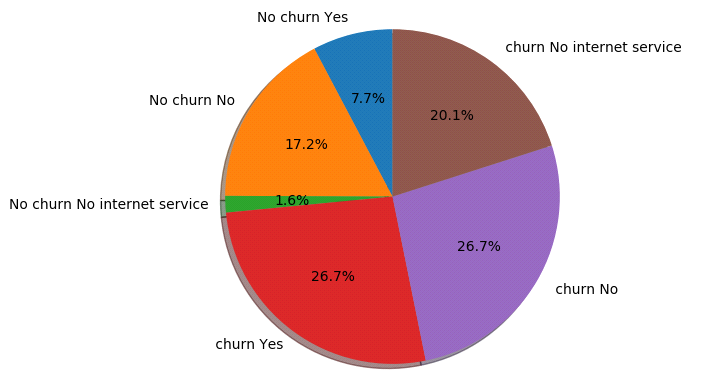
No internet service no churners : 1413

Yes churners : 545

No churners : 1211

No internet service churners : 113

* Customers without device protection churn more



Tech Support:

Yes no churners : 1734

No no churners : 2027

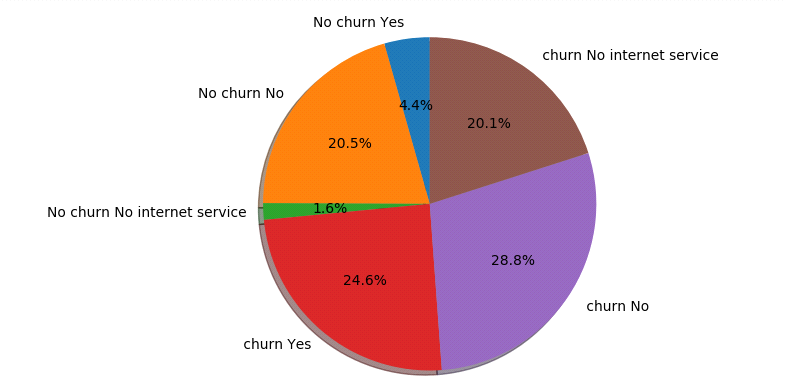
No internet service no churners : 1413

Yes churners : 310

No churners : 1446

No internet service churners : 113

* Customers without tech support churn more



Streaming TV :

Yes no churners : 1893

No no churners : 1868

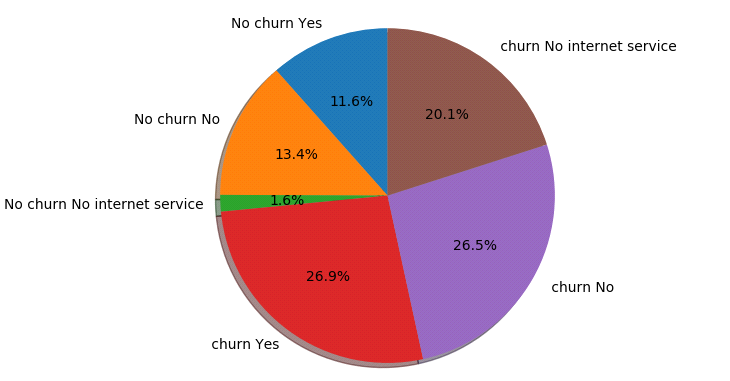
No internet service no churners : 1413

Yes churners : 814

No churners : 942

No internet service churners : 113

* Streaming TV has no influence over the results.



Streaming Movies :

Yes no churners : 1914

No no churners : 1847

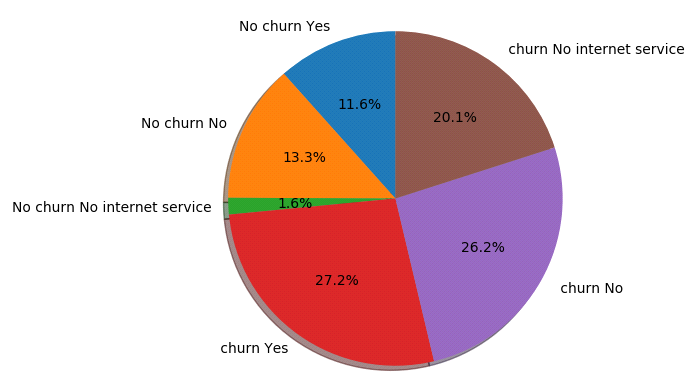
No internet service no churners : 1413

Yes churners : 818

No churners : 938

No internet service churners : 113

* Streaming movies has no influence over the results.



Contract :

Month-to-month no churners : 2220

Two year no churners : 1647

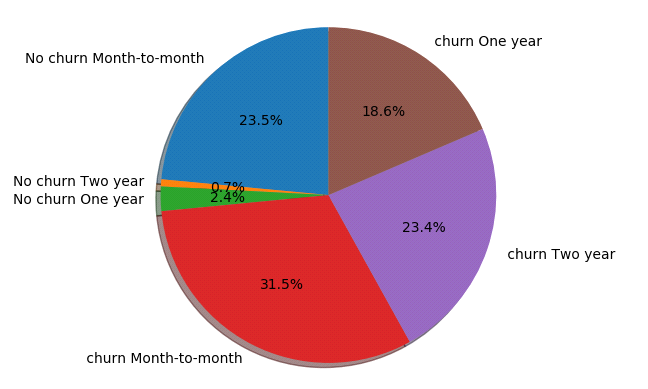
One year no churners : 1307

Month-to-month churners : 1655

Two year churners : 48

One yearchurners : 166

* Customers with month to month contracts churn more



Payment Method :

Mailed check no churners : 1304

Electronic check no churners : 1294

Bank transfer (automatic) no churners : 1286

Credit card (automatic) no churners : 1290

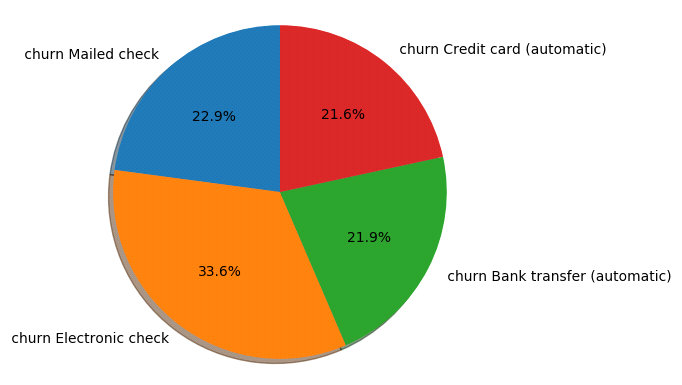
Mailed check churners : 308

Electronic check churners : 1071

Bank transfer (automatic) churners : 258

Credit card (automatic) churners : 232

* Customers who pay with electronic checks churn more

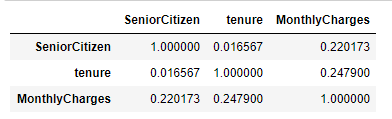


4- Data cleaning

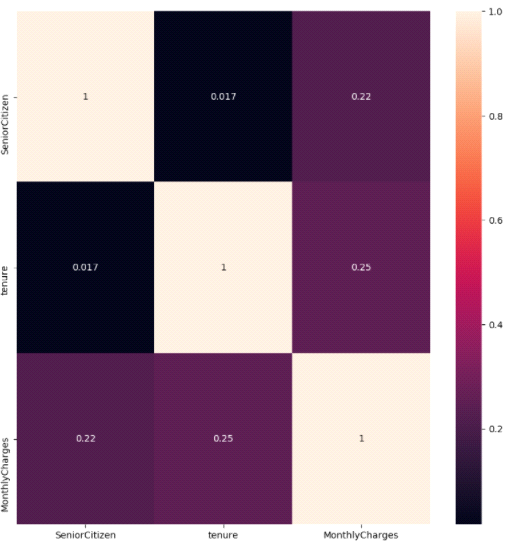
We dropped the columns CustomerID and Total charges in order to avoid overfitting that may lead to an inaccurate result.

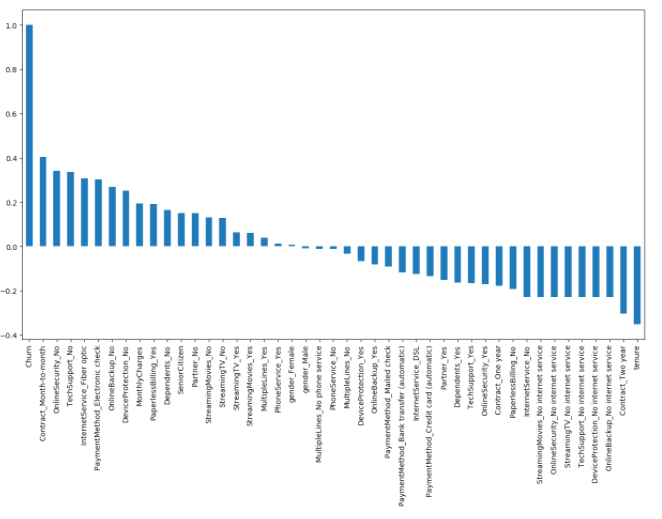
5- Data correlation

We clarified the dependencies between the three columns senior citizen, tenure and monthly chargers.



These graphs show how much are these columns dependent on each other and the relation between them.

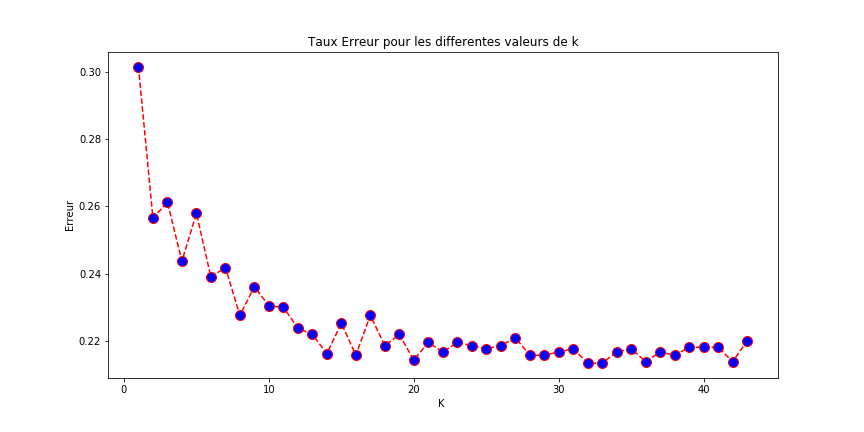




5-SFS/SBS Normalization:

Throughout this project we are going to apply different machine learning algorithms and find the accuracy of each.

**6-KNN Classification**

We don't know what is the optimal K value that gives maximum accuracy. So, we can write a for loop that iterates for example 45 times and gives the accuracy at each iteration. So that we can find the optimal K value.

**According to this graph, the best k value possible is 20 since it has the least error percentage.**

**-GridsearchCV**

We used this function to come up with the best possible value for the parameter K.

In this case K=29.

After applying the KNN method, the result was:

Accuracy of K-NN classifier on training set: 0.80

Accuracy of K-NN classifier on test set: 0.78

**-Confusion Matrix**

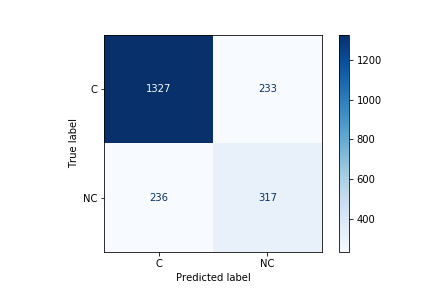
For logistic regression confusion matrix;

TN = 317

FP = 223

FN = 236

TP = 1327



**-classification report:**

Acuuracy should not be used as solely metric for imbalance datasets. There are some other metrics named as recall and precision.

Sometimes we get high recall and low precision or vice versa. There is another metric that combines both precision and recall like below. We will use F1 score to identify the best algorithm score.

precision recall f1-score support

0 0.85 0.85 0.85 1560

1 0.58 0.57 0.57 553

accuracy 0.78 2113

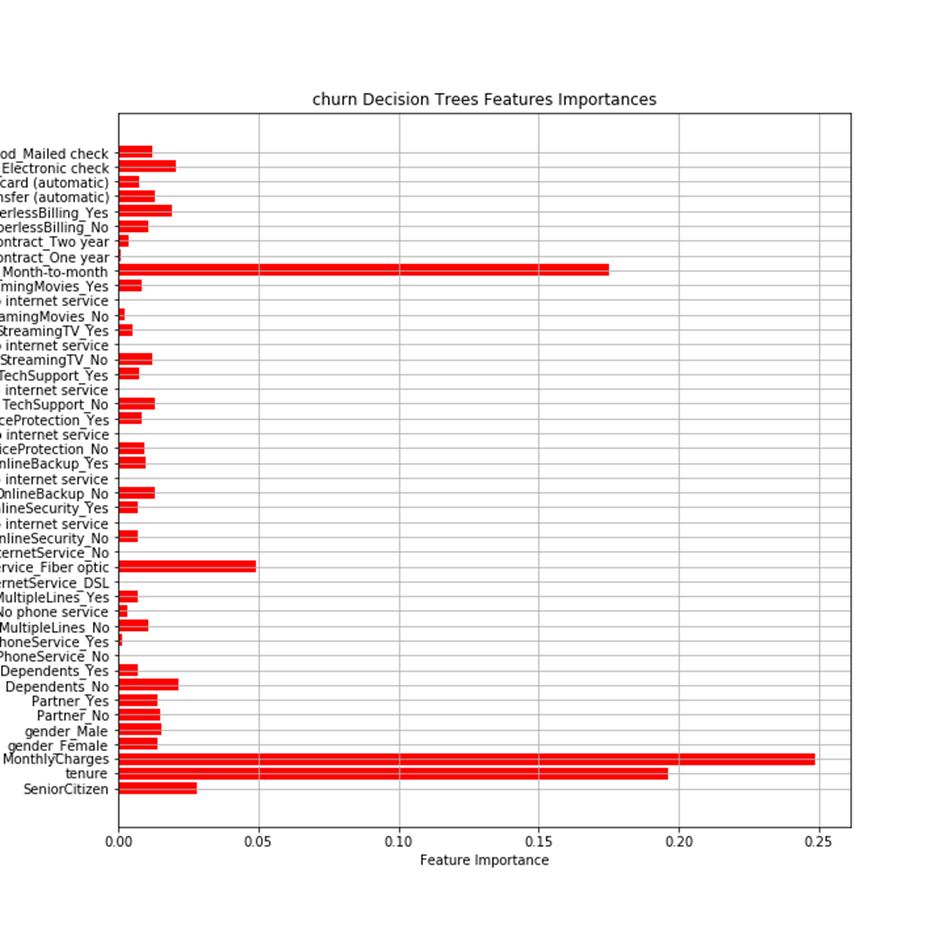
macro avg 0.71 0.71 0.71 2113

weighted avg 0.78 0.78 0.78 2113

**-DecisionTreeClassifier:**

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value.

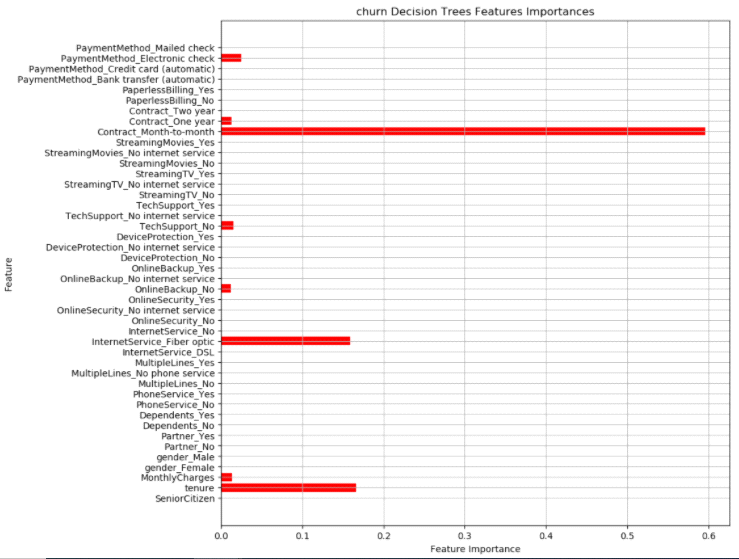
**1) All features:**

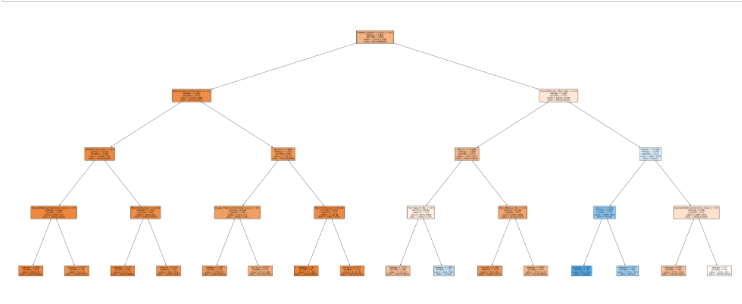


**1) After GridSearsch:**

churn\_feature\_importances :

we used the churn feature importance function in order to deduce the degree of impotance of each feature, consequently that helps us choose and found our work on correct basis





param grid helps us identify the most accurate method to follow. the result indicates that "entropy" is more accurate than genie.

{'criterion': 'entropy', 'max\_depth': 4}

final model gave us the following results :

train score = 0.7975659229208925

test\_score = 0.7898722195929957

**-Naive Bayes Classification:**

Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems. It is called Naive Bayes or idiot Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable.

Naive Bayes accuracy found is : 0.6942735447231424

**-SVM Classification :**

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. ... This is where the LSVM algorithm comes in to play.

SVM accuracy found is : 0.7983909133932797

**-Logistic regression classification :**

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

Logistic Regression accuracy found is : 0.8026502602934217

​

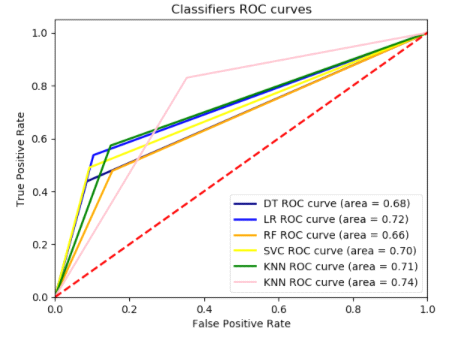
**-Random Forest Classification :**

It works in four steps:

* Select random samples from a given dataset.
* Construct a decision tree for each sample and get a prediction result from each decision tree.
* Perform a vote for each predicted result.
* Select the prediction result with the most votes as the final prediction.

Random Forest accuracy for 5 trees is : 0.7496450544249882

To determin the most accurate method we used the classifier ROC.



the graph indicates that most of the methods are similar and give us almost the same results