Information about the data

Link

https://www.kaggle.com/blastchar/telco-customer-churn

Context

"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

Content

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

Customers who left within the last month – the column is called Churn

Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges

Demographic info about customers – gender, age range, and if they have partners and dependents

Dataset Details

customerID - Customer ID

gender - Whether the customer is a male or a female

SeniorCitizen - Whether the customer is a senior citizen or not (1, 0)

Partner - Whether the customer has a partner or not (Yes, No)

Dependents - Whether the customer has dependents or not (Yes, No)

Tenure - Number of months the customer has stayed with the company

PhoneService - Whether the customer has a phone service or not (Yes, No)

MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)

InternetService - Customer’s internet service provider (DSL, Fiber optic, No)

OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)

OnlineBackup - Whether the customer has online backup or not (Yes, No, No internet service)

DeviceProtection - Whether the customer has device protection or not (Yes, No, No internet service)

TechSupport - Whether the customer has device protection or not (Yes, No, No internet service)

StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)

StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)

Contract - The contract term of the customer (Month-to-month, One year, Two year)

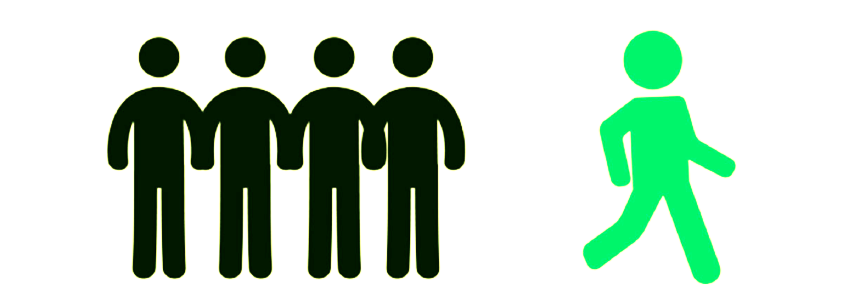
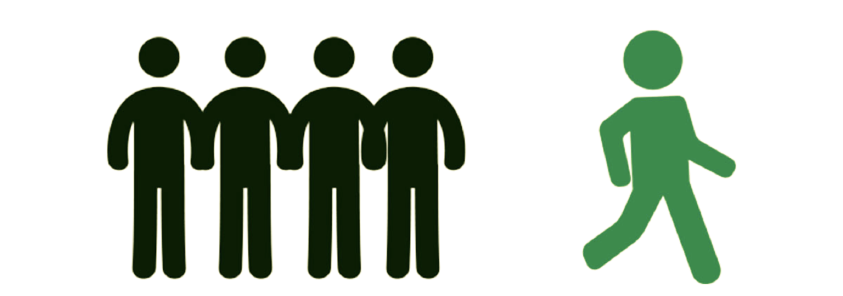
PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)

PaymentMethod - The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

MonthlyCharges - The amount charged to the customer monthly

TotalCharges - The total amount charged to the customer

Churn - Whether the customer churned or not (Yes or No)



**Purpose**

Nowadays, it’s very easy to change between companies to serve our needs, but for the company itself, the loss of a customer comes at a great cost. It is much more rewarding for the company to retain the current customers than to invest in acquiring new ones. This way, reducing customer churn has become a fundamental skill for a company.

Some approaches that have been practiced are based on identifying customers who are at high risk of churning and encourage them to stay. Thus, **the purpose of this analysis is to predict customer churn based on his characteristics.**

In this article, will be used a Telco dataset and go through the following steps to develop a Churn prediction model:

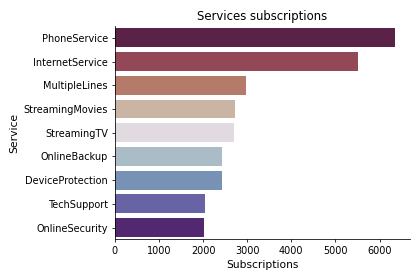
1. Exploratory data analysis
2. Feature engineering
3. Modeling
4. Results

Next it will be shown some analysis and results, but for more details you can check the [GitHub page](https://github.com/margaridacrsp/udacity/tree/master/Final_Project).

**Dataset**

The dataset has 7043 rows and 21 columns, where each row represents a customer and the columns his characteristics:

* Demographic information: gender, age range and if the customer has partners and/or dependents;
* Services information: phone, multiple lines, internet, online security, online backup, device protection, tech support and streaming TV and movies;



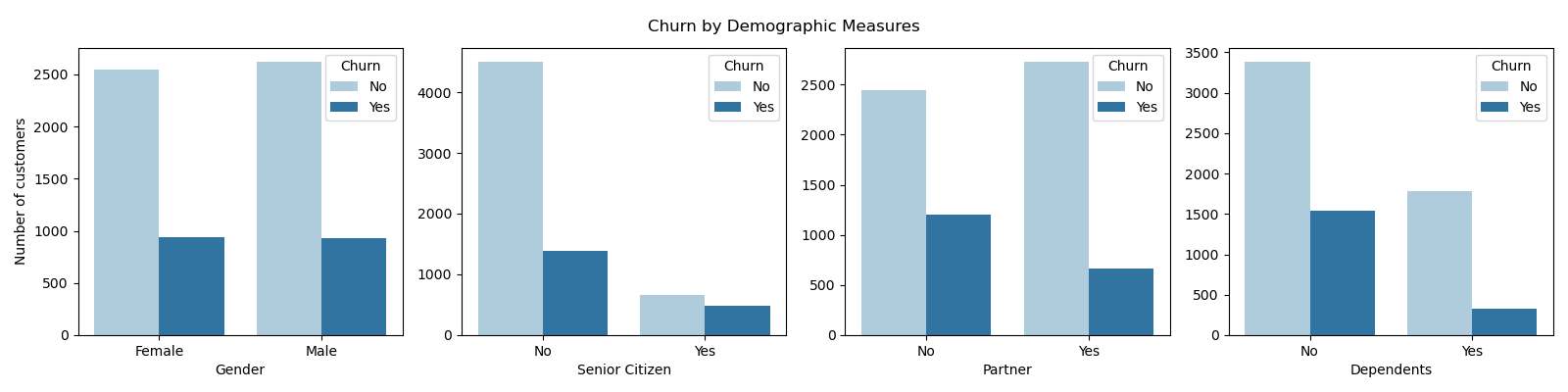
* Account information: longevity as customer, type of contract, payment method, paperless billing, monthly charges ant total charges;
* Churn information: if the customer has left the company in the last month.

**1. Exploratory data analysis**

At this stage, it is important to understand the relationship between the various variables and the target variable, to gain some sensitivity to its impact and weight on the final forecast. Let’s see some examples:

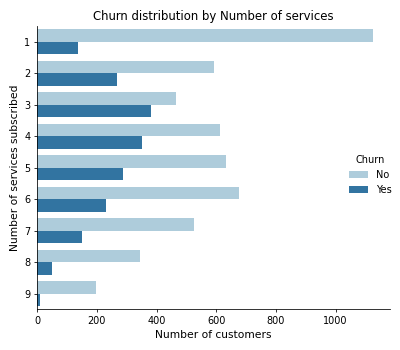
**· Churn by demographic variables**

The proportion of churns is very similar for both genders (~26%), however for the remaining demographic variables there is a significant difference between them: 41,7% for senior customers and 23,6% for non-senior customers; 19,7% for customers with partners and 33% for customers without partners; and customers with dependents have 15.5% while customers without dependents have 31.3% of churn.



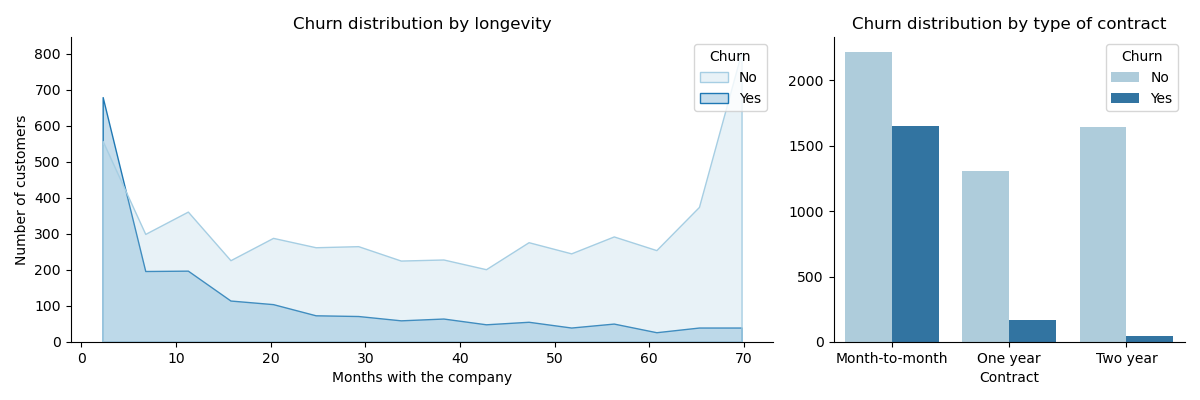
**· Churn by services variables**

In regards to number of services subscribed, the highest percentage of churn customer is between the customers that had 3, 4 or 5 services subscribed.

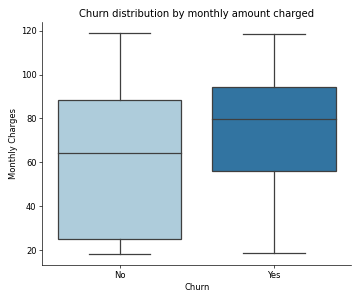


**· Churn by account variables**

The first graph shows the longevity of customers, where it is clearly possible to see that most churn customers remain in the company for only a few months. This makes sense with the second graph, where it can be seen that the churn customers opted mainly for monthly contracts.



Regarding contractual charges, churn customers pay on average more than other customers, per month. This may be one of the reasons why the customer ceases his relationship with a company.

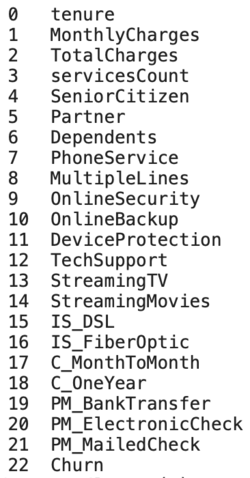


**2. Feature engineering**

In this section, the raw variables were transform to create the features that looked promising to train the model on. The strategy was the following:

* Numerical variables remained numerical;
* Categorical variables were transformed in binary, applying the Dummy Variables method.

The final model ended up with 22 new features.



**3. Modeling**

To train the model and predict new results the following steps were made:

1. Separate the features variables from the target variable;
2. Split the data into training and test sets, with 67% train and 33% test;
3. Create pipeline with the scaler and the classifier. For the scaler, was chosen the standard scaler, that standardize features by removing the mean and scaling to unit variance and, for classifier, were chosen 5:  
    a. Random Forest  
    b. Gradient-Boosted Tree  
    c. Decision Tree  
    d. K-Neighbors   
    e. Logistic Regression
4. Define the performance metrics to compare the model:  
    a. Accuracy  
    b. Area under the curve (AUC)  
    c. F1-score  
    d. Time to train the model

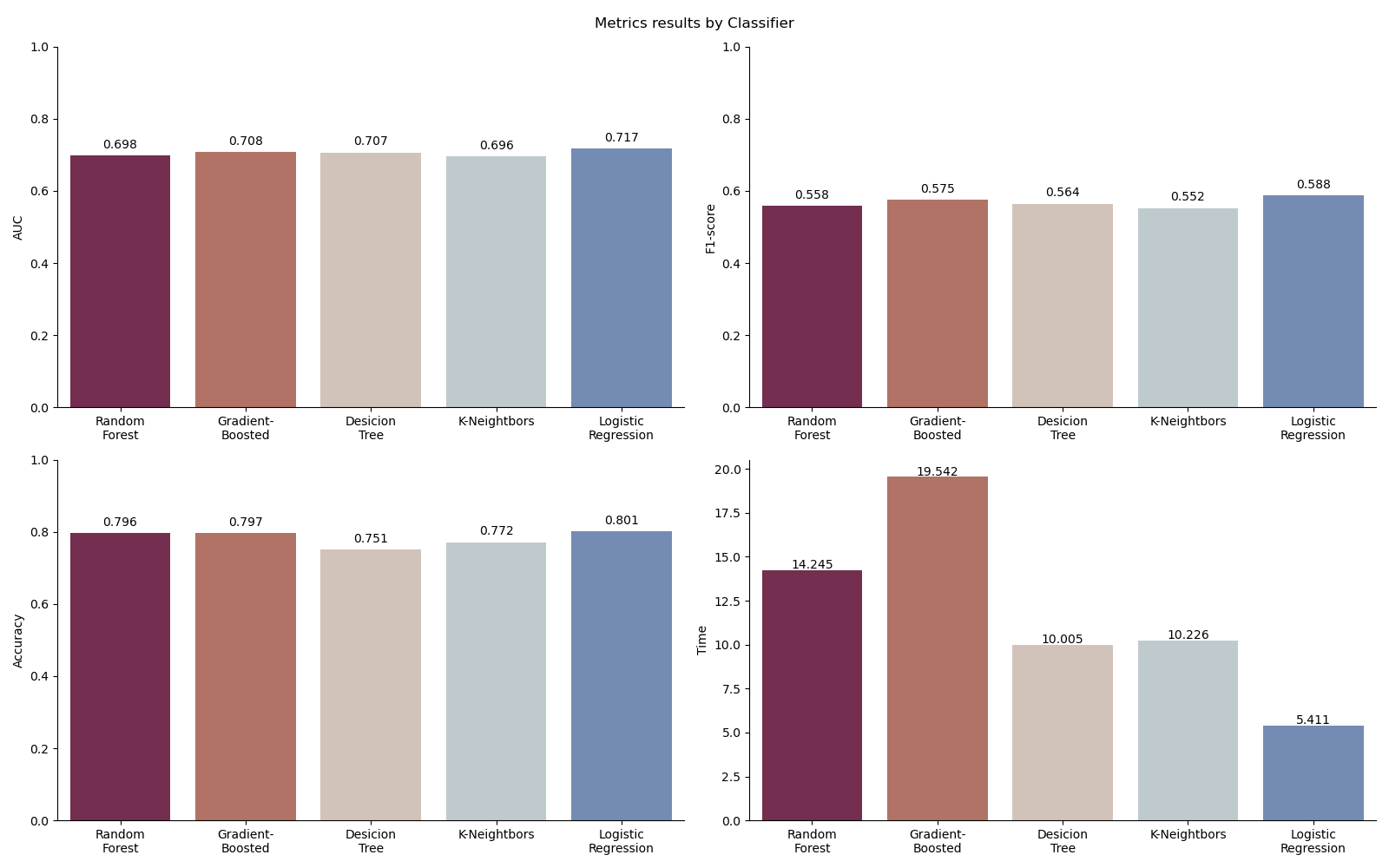
The best model should optimize the AUC metric. This metric was chosen has winning because our target variable is binary, so it’s a binary classification problem and AUC is a good way for evaluation for this type of problems.

5. Train the model and tuning hyper-parameters with grid search method

**4. Results**

As can be seen in the graphs below, the **Logistic Regression** model obtained the best results in the 4 metrics: **area under curve equal to 0.717, 0.588 of f1-score and accuracy equal to 0.801 in 5.4 seconds of execution**.

However, with the exception of the execution times, the values of the other three metrics are very identical among the 5 analyzed classifiers.

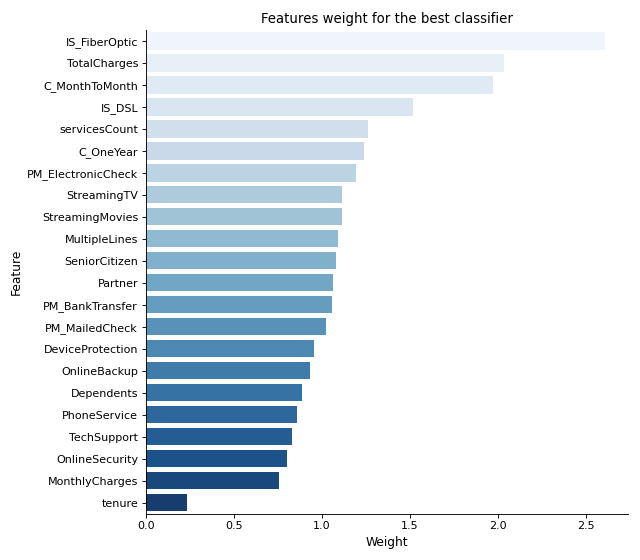


Finally, the importance of each feature variable was analyzed in order to understand the impact of each of them on the customer’s churn.

The first most relevant resource is the fiber optic internet. This variable indicates whether the customer has internet via optical fiber, but impacts not only the internet service as also in the television service. So, in a Telco company perspective, this service has a huge impact on the way the customer sees the company, meaning that if the internet and tv signal is bad the customer will be dissatisfied.

The second most relevant feature is the customer total charges. This variables is connected to the time the customer with the time that the customer is in the company, but also with the monthly fees paid to the company.

The third most relevant feature is the Month-to-Month contract and this one is a easy one, because with a possibility of staying with the company must longer with the other two types of contracts, if the customer chooses this one, is a good indicator that the customer doesn’t have intentions to stay in the company for a long time. This variable can be seen as an engagement metric.



**Conclusions**

The objective of this project was to predict the churn of customers based on demographic, services and account data.

Some exploratory data analysis was done to understand which variables could have an impact on the churn event. After exploring the data, the columns of interest were transformed into numerical or binary columns to be used in the modeling phase.

The dataset was divided into training and testing, in the proportion 67% and 33%, respectively, and 5 classification models were trained and the hyper-parameters where adjusted using Grid Search.

Finally, the model that showed the best results for predicting churn was the Logistic Regression, with an AUC of 0.717. This model obtained the highest scores in all 4 metrics chosen.

**Future work**

* Work with a larger data set to train the model on a larger sample;
* Explore other models, such as Support Vector Classification (SVC);
* Treat numeric columns, such as monthly charges and total charges, in another way, for example grouping the several values.

**References**

Dataset  
<https://www.kaggle.com/blastchar/telco-customer-churn>