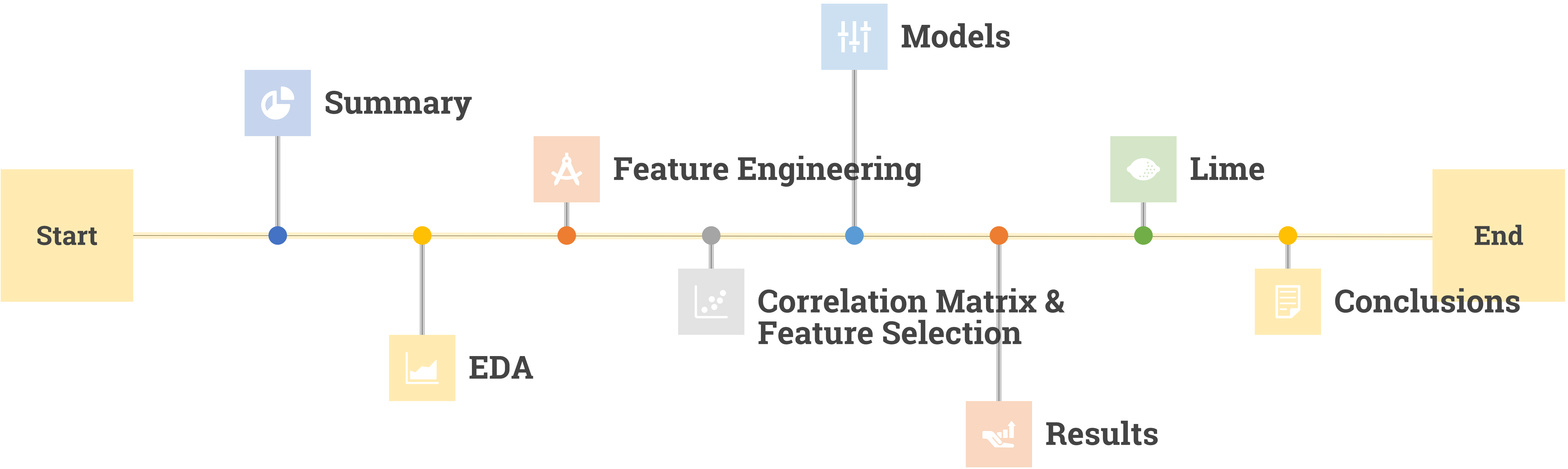
Applied Machine Learning for Churn Rate Prediction

Telco application

Gateano Costa, Andrea Huscher,
Federico Porcu, Niccolò Salvini

Agenda


Start
Summary
EDA
Feature Engineering
Correlation Matrix &
Feature Selection
Models
Results
Lime
Conclusions
End

Section 1

Summary

Project Summary:

1 **Stakeholder:**

Telco company

2 **Objectives:**

Predict the churning rate.

Identify which variables are signals of a possible client churning.

3 **Methodology:**

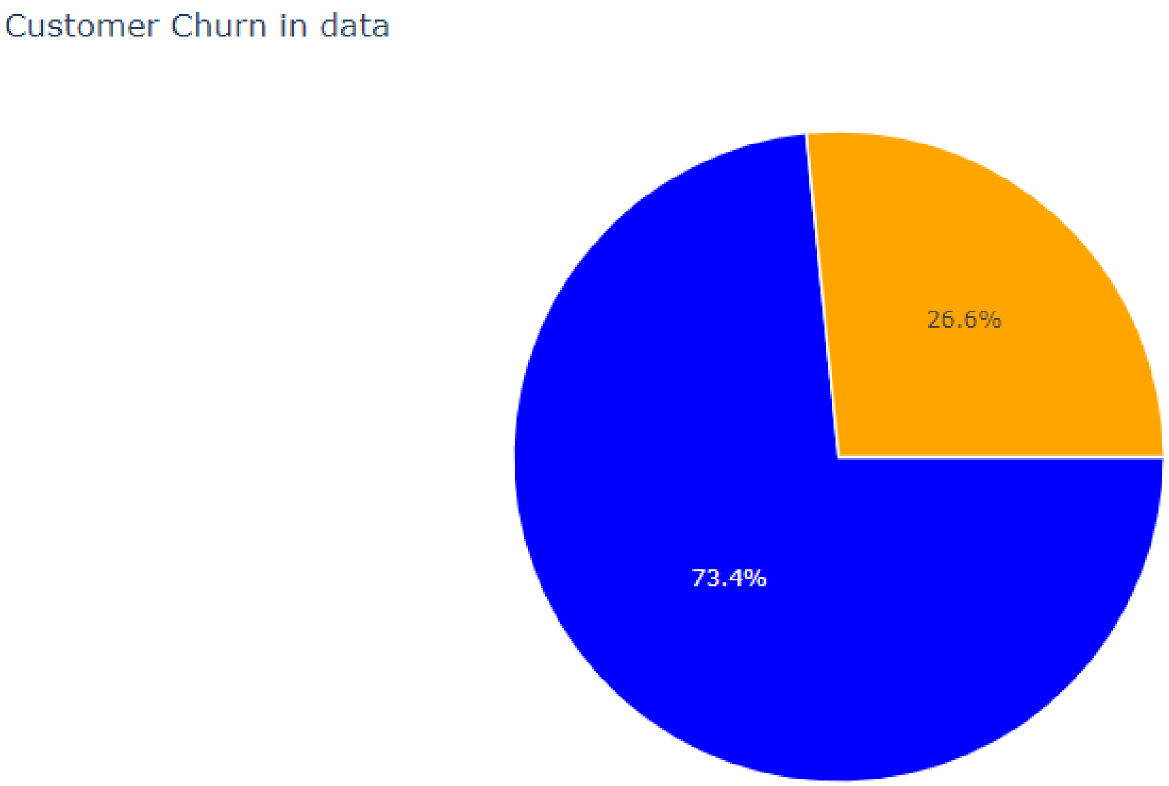
Building five different classification models in order to predict the churning rate:

- Logistic Regression
- Support Vector Classifier
- Random Forest
- Gradient Boosting
- XGBOOST

4 **Evaluation:**

We will see that the dataset is unbalanced, so as evaluation metrics we used mainly the F1-score.

Dataset Overview



- 7043 observations.
- 21 variables: 16 categorical, 3 numerical, 1 object, 1 target.
- It stores various data for each customer of a telecommunication company.
- Problem: the dataset is unbalanced. Here is displayed the target distribution.
- There are only 11 missing values, we decided to drop them (please find the reasoning in the notebook).

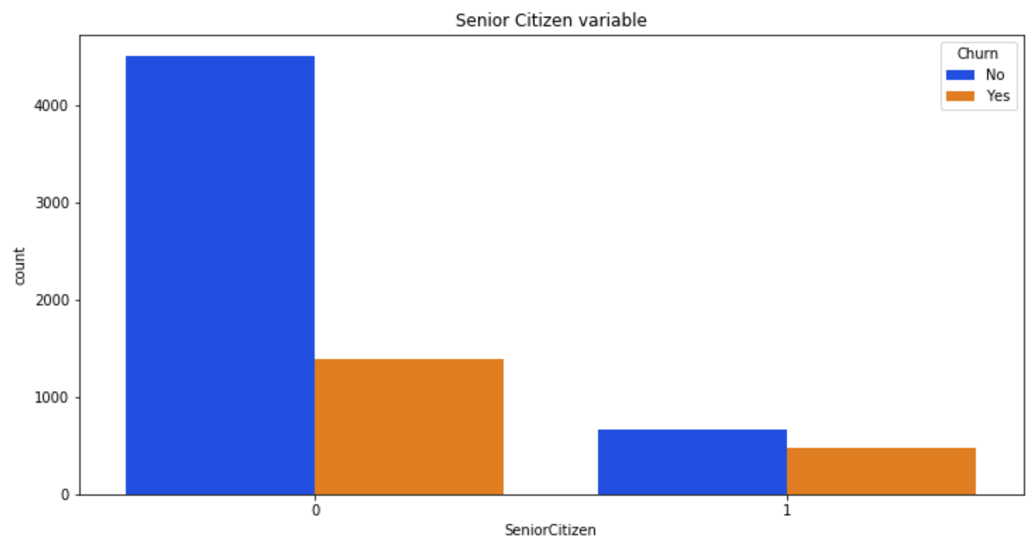
Section 2

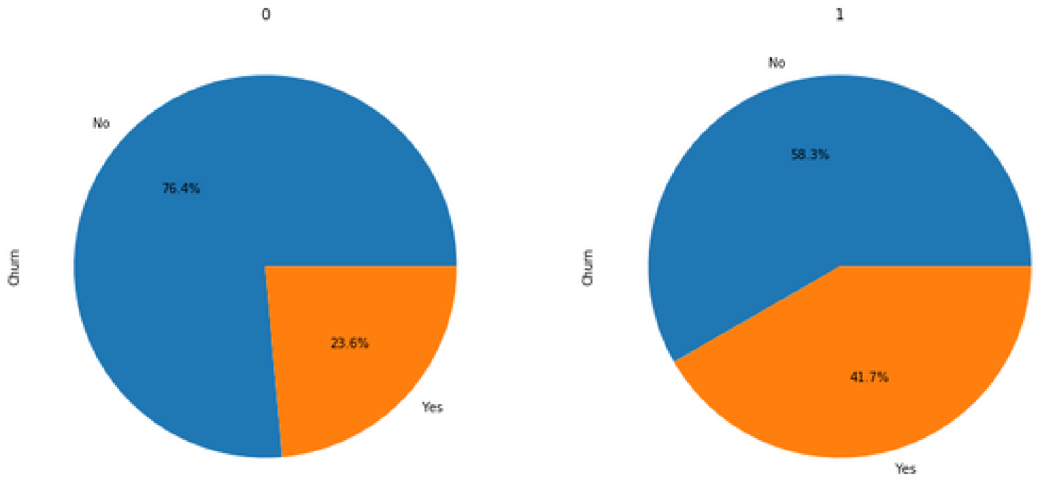
EDA

EDA:

1. Significant Categorical Variables
2. Numerical variables

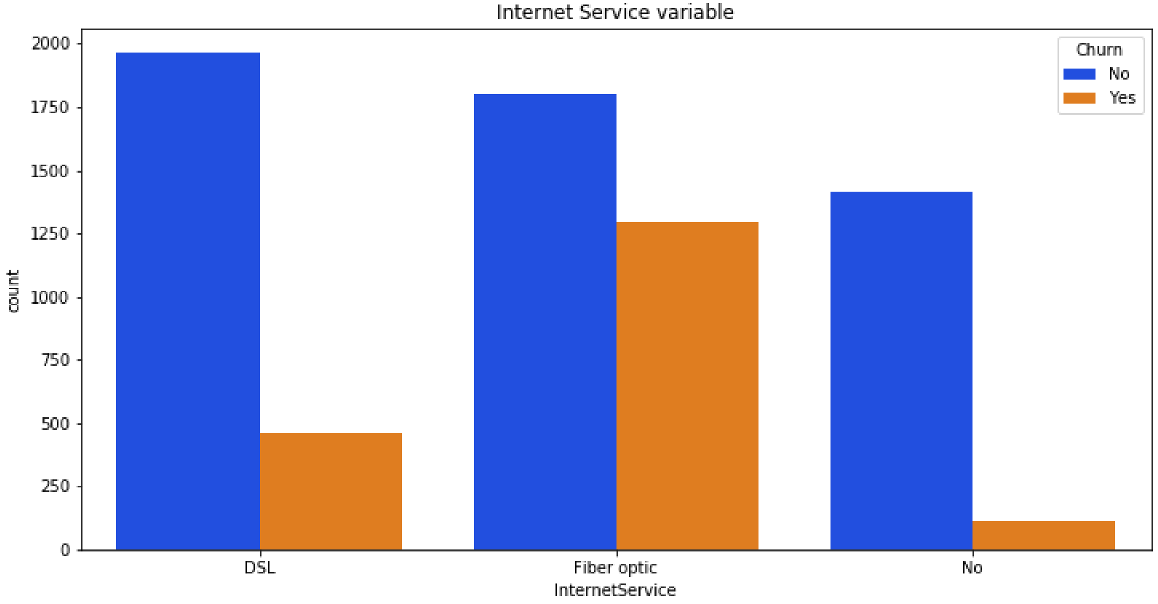
Significant Categorical Variables

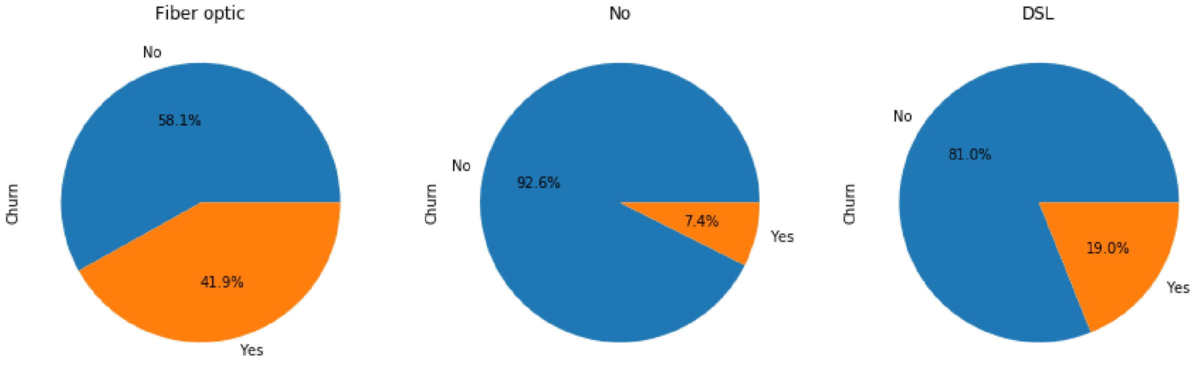




Senior Citizen

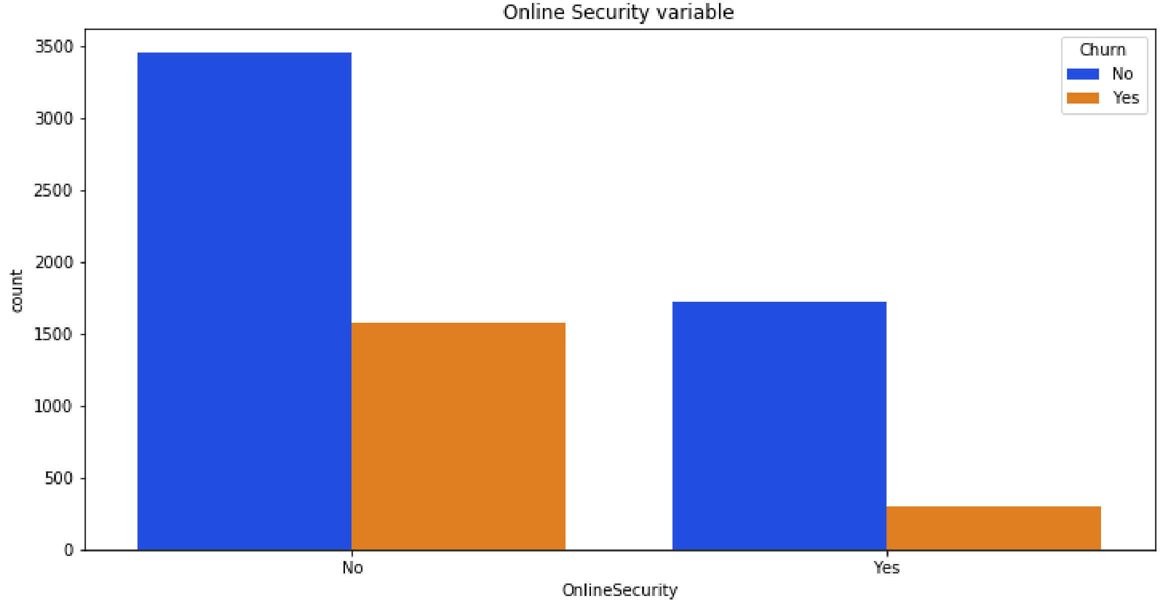
- This variable states if a client is Senior or not.
- Here we can see that this company has less senior clients.
- But the churning proportions are very different. In fact in senior clients the churning rate is much higher than the one of the others, 42% against 24%.

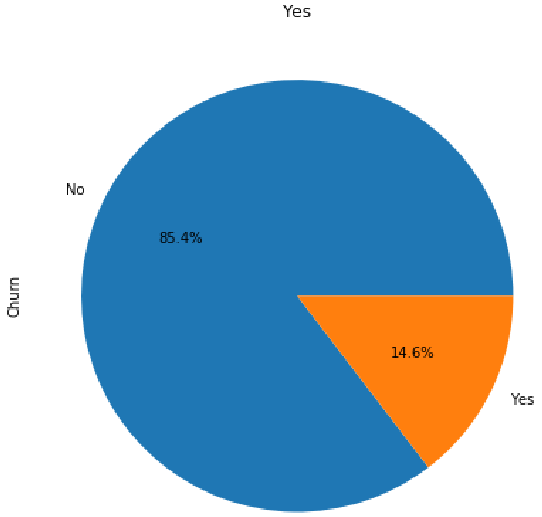


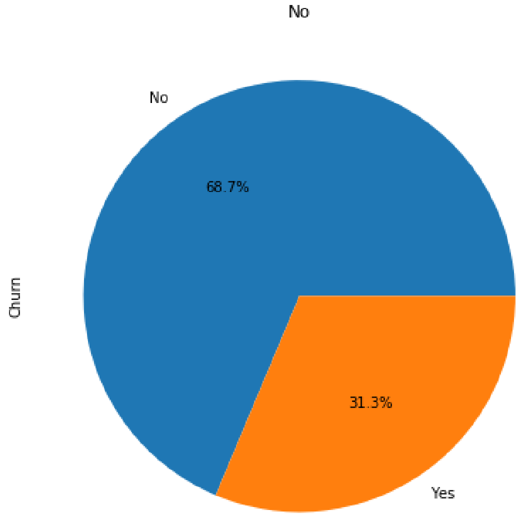


Internet Service:

- This variable states if a client has internet in his subscription, and if he does have internet which kind of connection he owns: fiber optic or DSL.
- We can see that the majority of the clients has fiber optic connection. Nowadays this makes sense, because of the technological advancement of this technology.
- Another important aspect is that the fiber optic class has the highest churning rate: about 42%.

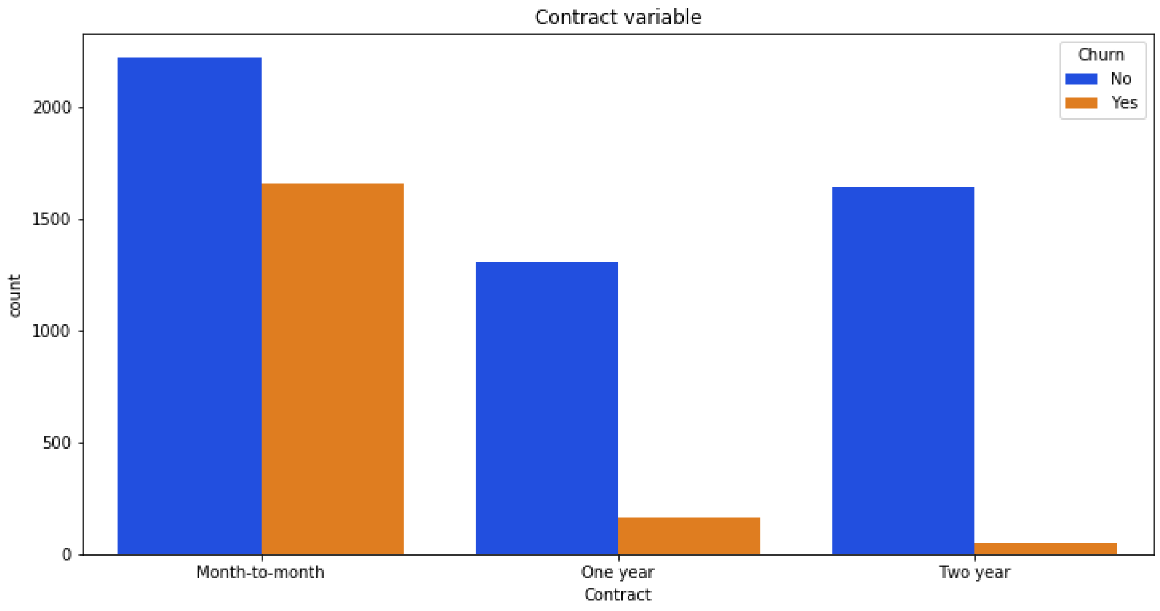


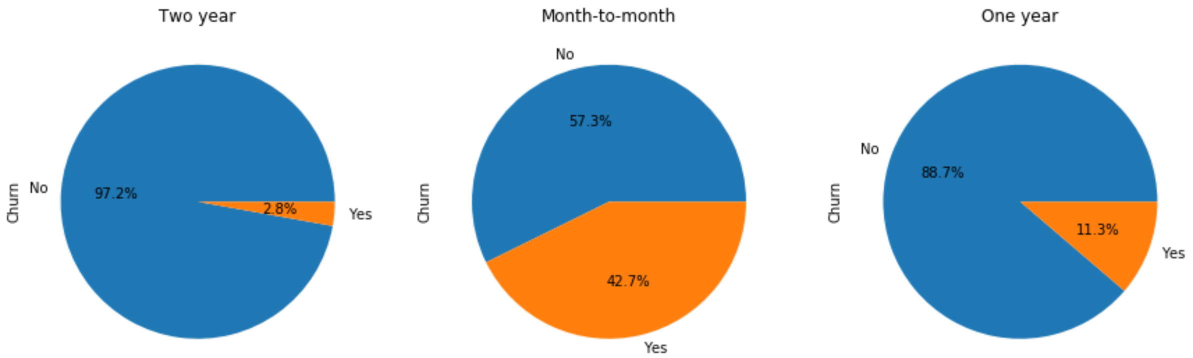




Online Security:

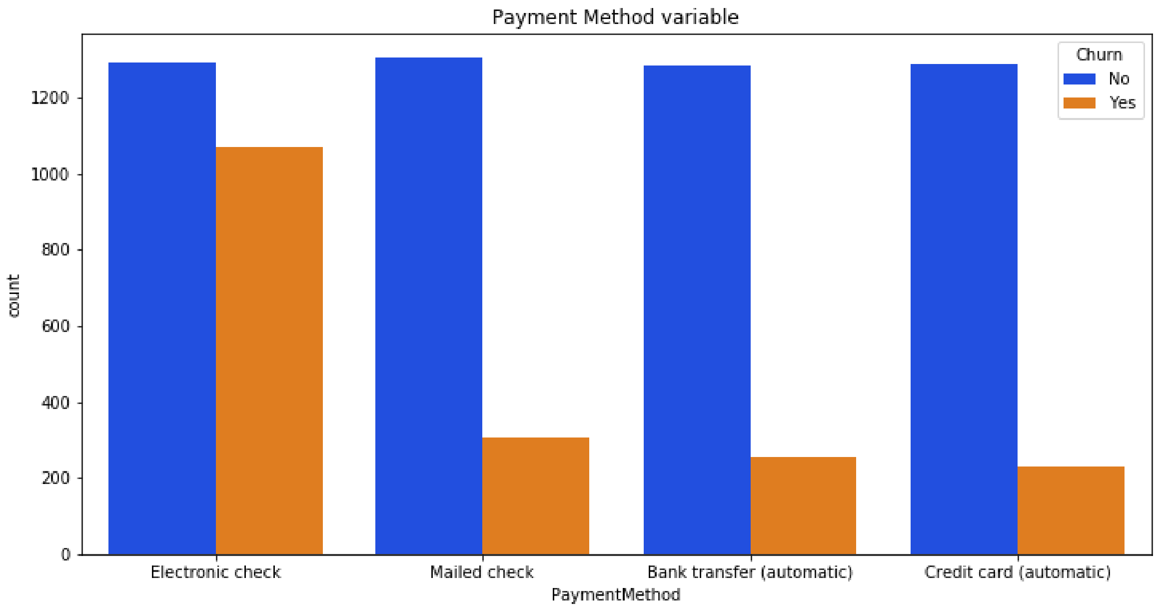
- This variable tells if a client has or not the online security service.
- Clients that don't have this kind of service have higher churning rate: about 31%.

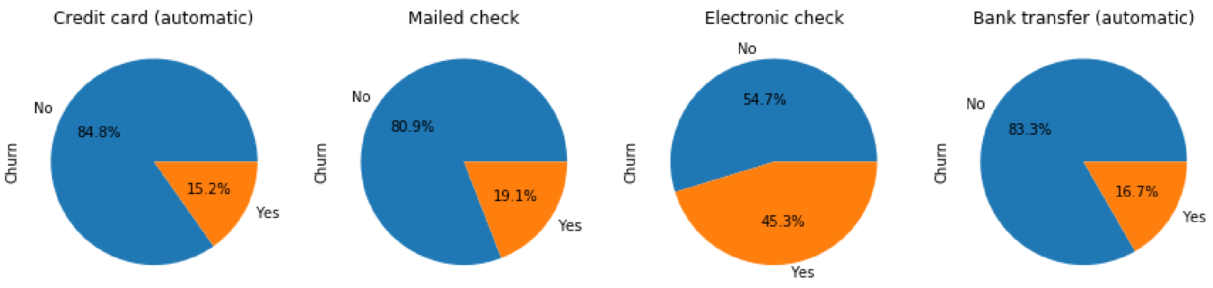




Contract:

- Important variable: tells which kind of billing contract each client has.
- We see that the majority of clients are billed monthly.
- We see also that the month to month class is the one with the highest churning rate: 43%.
- This variable can be seen as an engagement index: the longer the contract the more loyal is the client, thus less risky to churn





Payment Method:

- This variable tells which kind of method of payment each client has.
- The payment method with the highest churning rate is electronic check: 45%
- Maybe it's due to the fact that with this method it is easier for the client to quit the contract.

Other Categorical variables:

Gender

Partner

Dependents

Phone Service

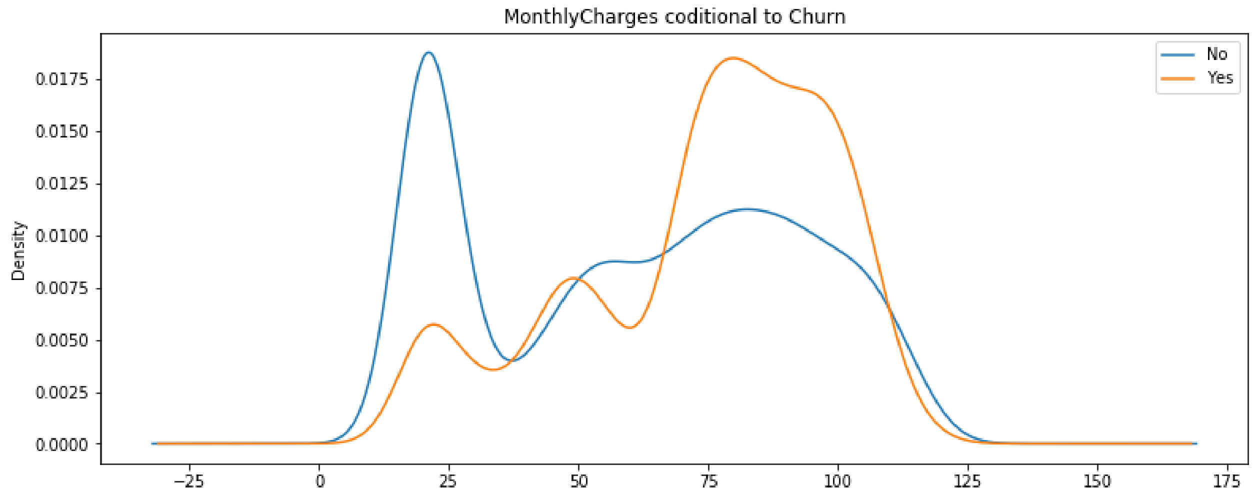
Streaming Movies

Streaming Tv

Multiple Lines

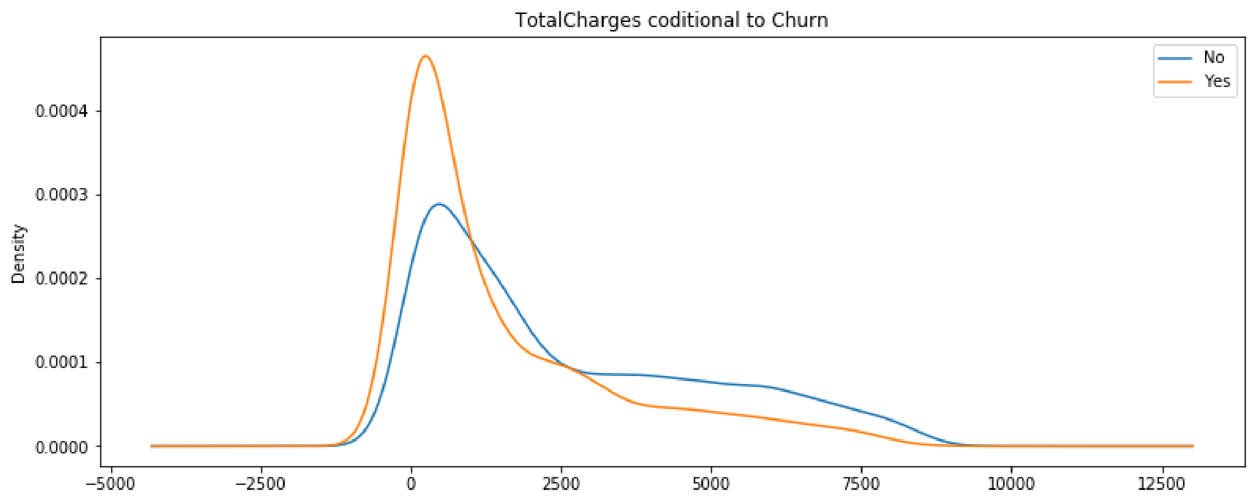
Numerical Variables

Monthly Charges



- **The Two Distributions of MonthlyCharges conditioned to the two classes of Churn are quite different.**
- **Customers who pays more monthly are more likely to churn while the opposite is true for customers who pay less.**

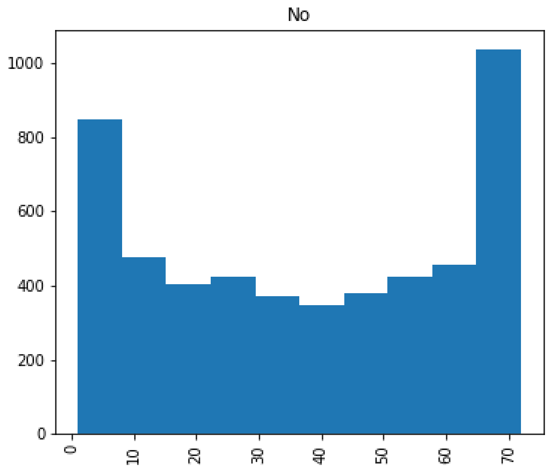
Total Charges

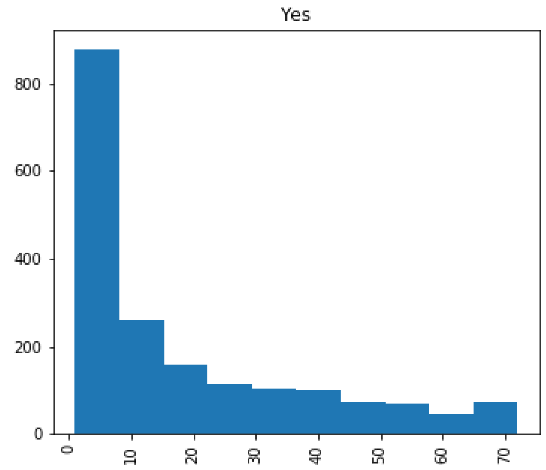


- The **Two distributions** of TotalCharges conditioned to the two classes of Churn **are very similar**.

- Customers who do not churn appear to have a Total Charges of money spread on a wider range, while the churning are more clustered within the lower values.

Tenure

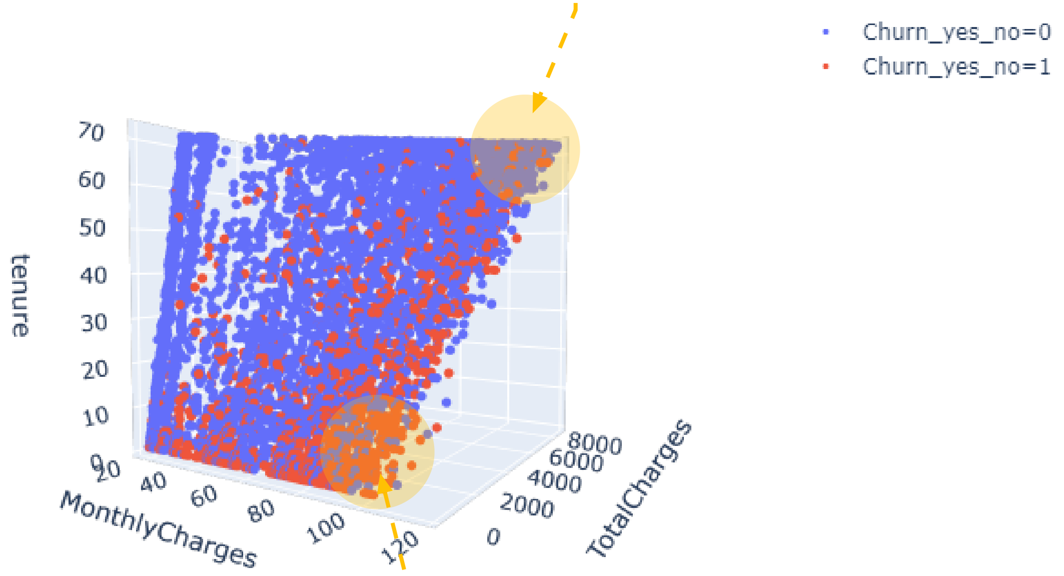




- The two histograms of tenure conditioned to the two classes of Churn are really different.
- The tenure of customers who churns are more likely to be short: this means that lots of customers decided to churn fast.
- This orientation seems to imply a good newly entering customers contract promotion.

3D Plot Flagged for Churn

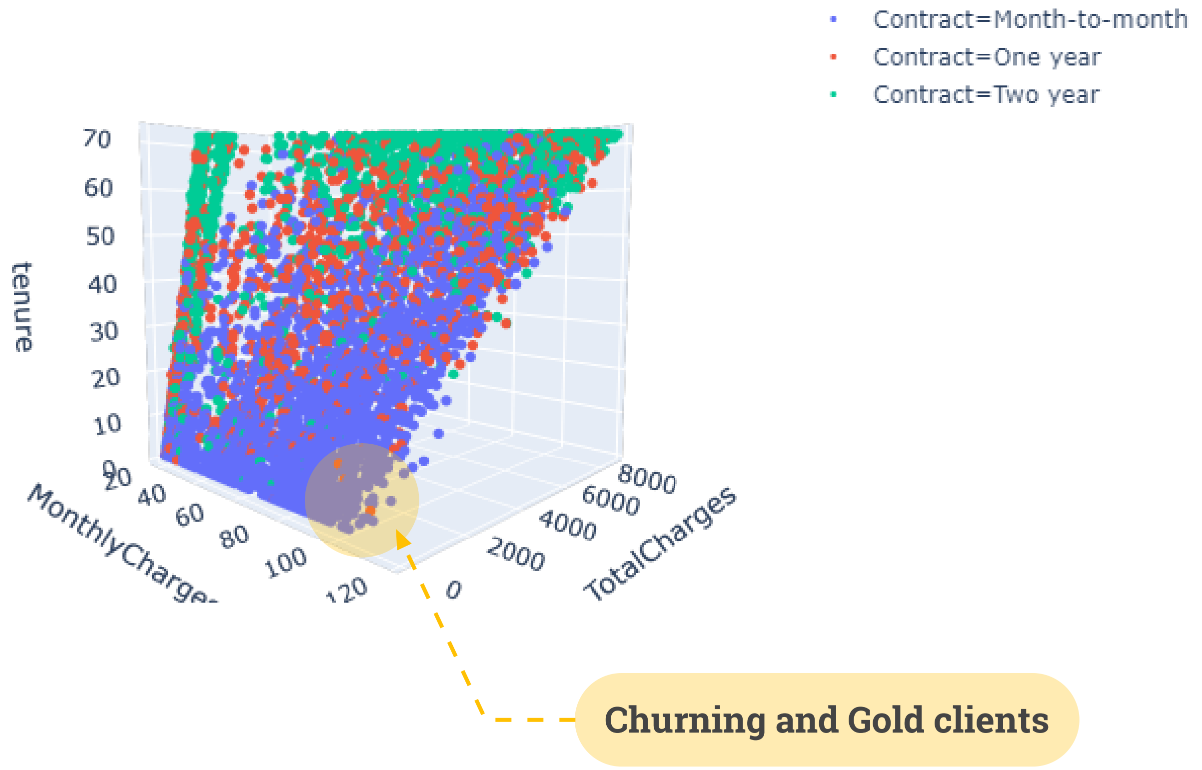
Area of Interest
2nd Goal



Area of Interest
1st Goal

- The clients that don't churn are equally distributed on the space.
- The clients which churned are mainly grouped where the tenure is low and where the MonthlyCharges is higher.

3D Plot Flagged for Contract



- This variable can potentially make the difference. It horizontally divides customers. Month two Month people generally have a low tenure because they test the service and then evaluate the churning option.
- But they are also sparser than the other two type of contract. It seems that this variable could be very promising for the classification of churn.
- The reason could be related to the fact that people with Month-to-Month contracts are less engaged that those who have different types of contract.

Section 3

Feature Engineering

Variables Created

- ## Number of Services

  A **numeric** variable that explores the number of combined services that a customer has activated.

- ## Tenure binned

  The transformation of a numerical variable into a **categorical** one either according to a convenient economic choice and a statistical consideration.

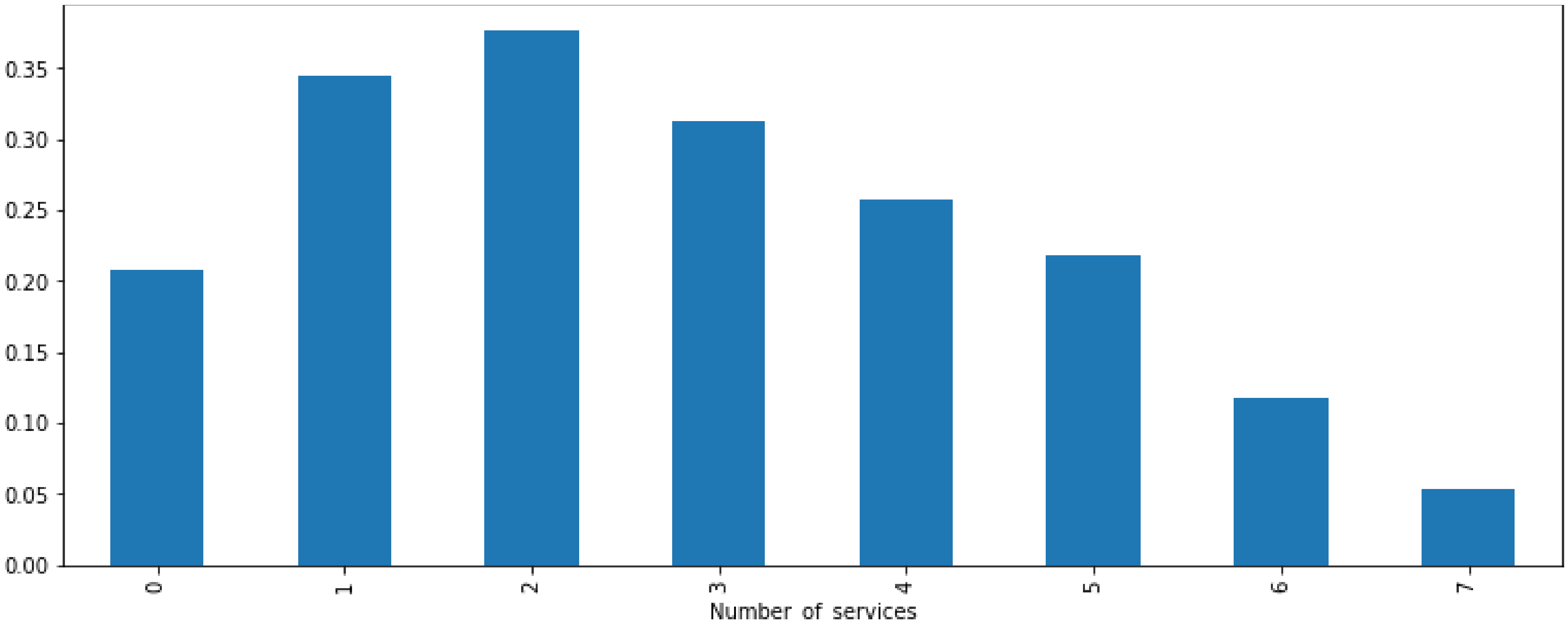
- ## MonthlyCharges binned

  The transformation of MonthlyCharges into a **categorical** variable that contains ranges of monthly expenditures.

- ## Exp_vs_Real

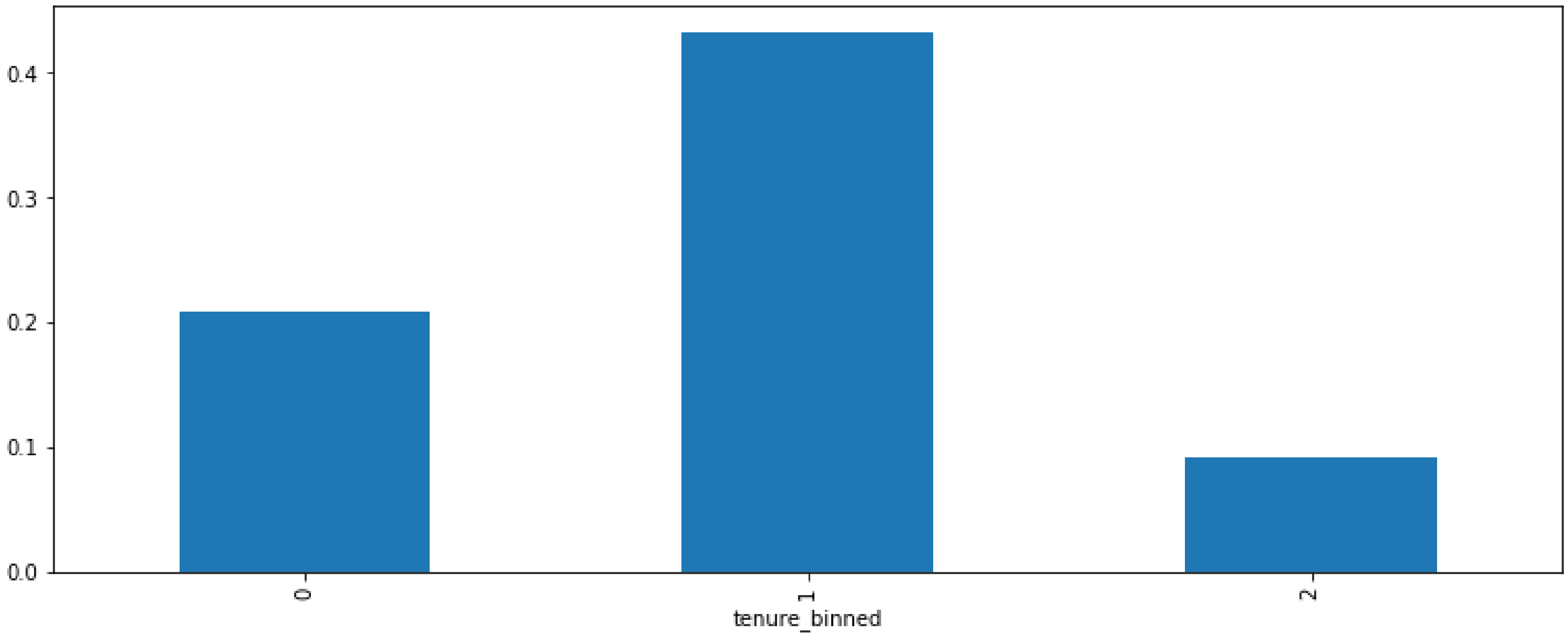
  A **numeric** variable that has the purpose to demonstrate the difference between what the customer expected to paid and what was really the contract bill.

Number of Services Churn: yes proportion



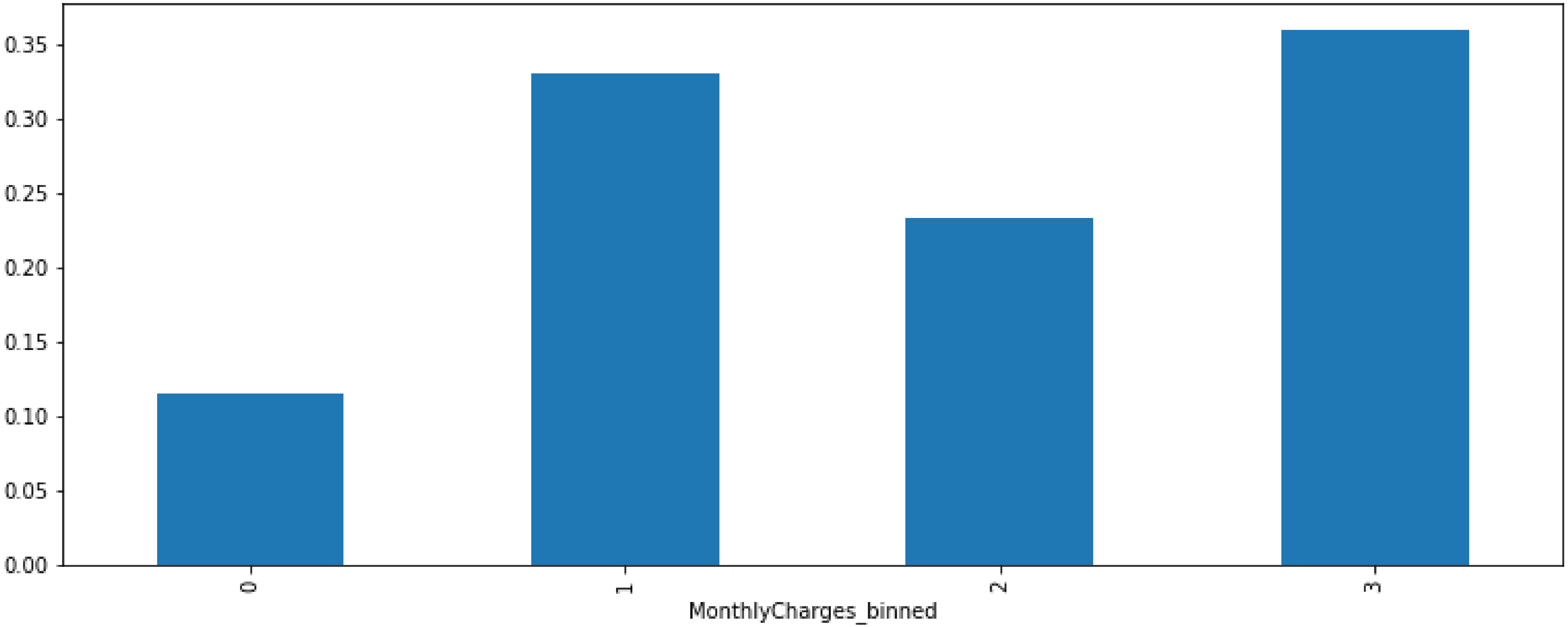
Customers with different number of services Churns with different probabilities. Customers with more services churn with lower probability.

Tenure binned Churn: yes proportion



The Churn rate is different for the 3 classes of tenure.

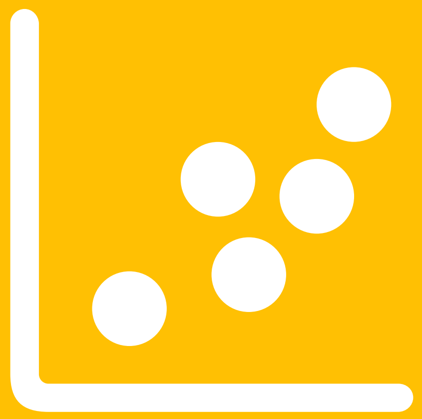
Monthly Charges binned Churn: yes proportion



The Churn rate is different for the 4 classes of monthly charges.

Section 4

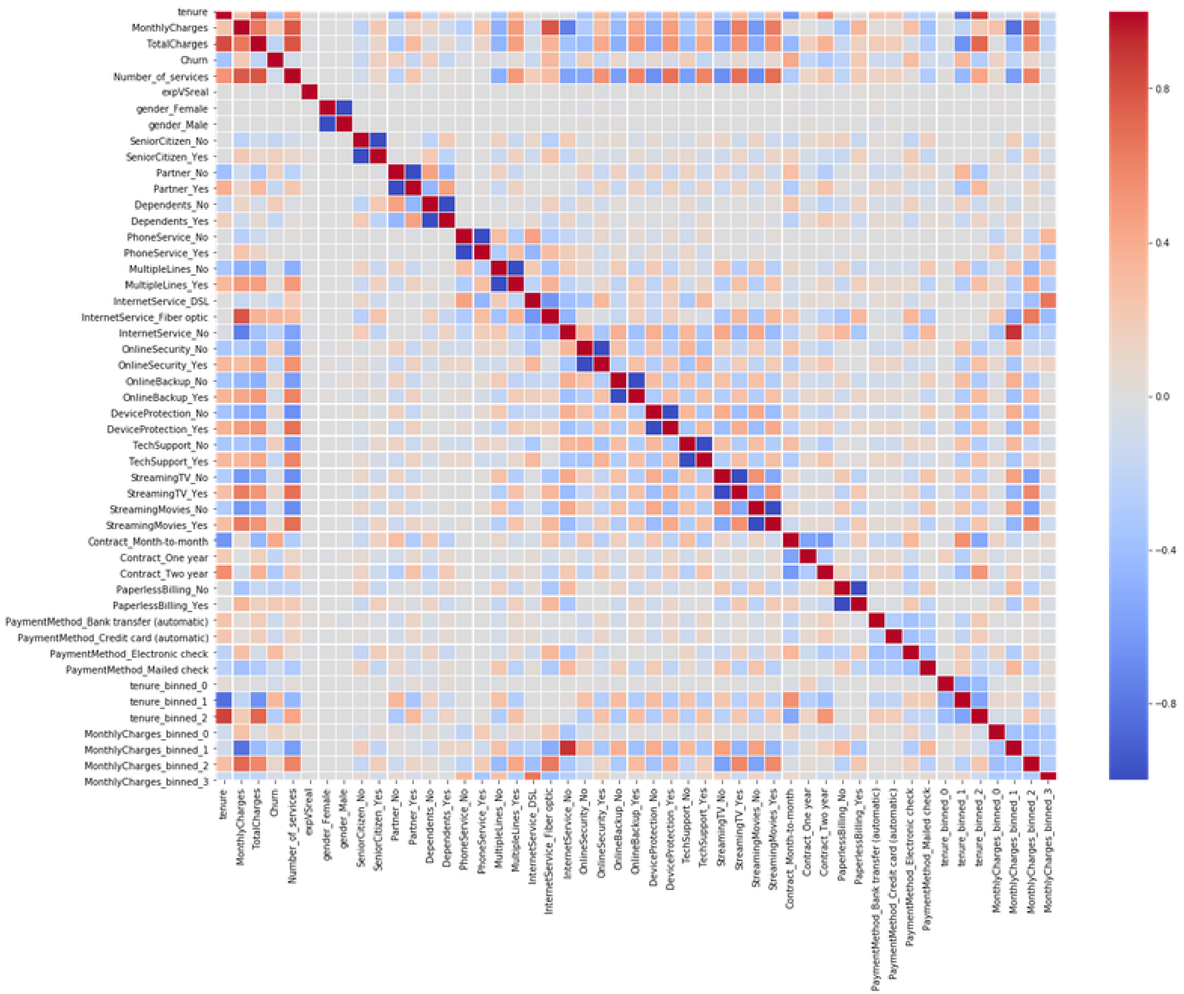
Correlation Matrix and Feature Selection



From now on we will follow Two approaches:

- Mean encoding
- One-Hot encoding

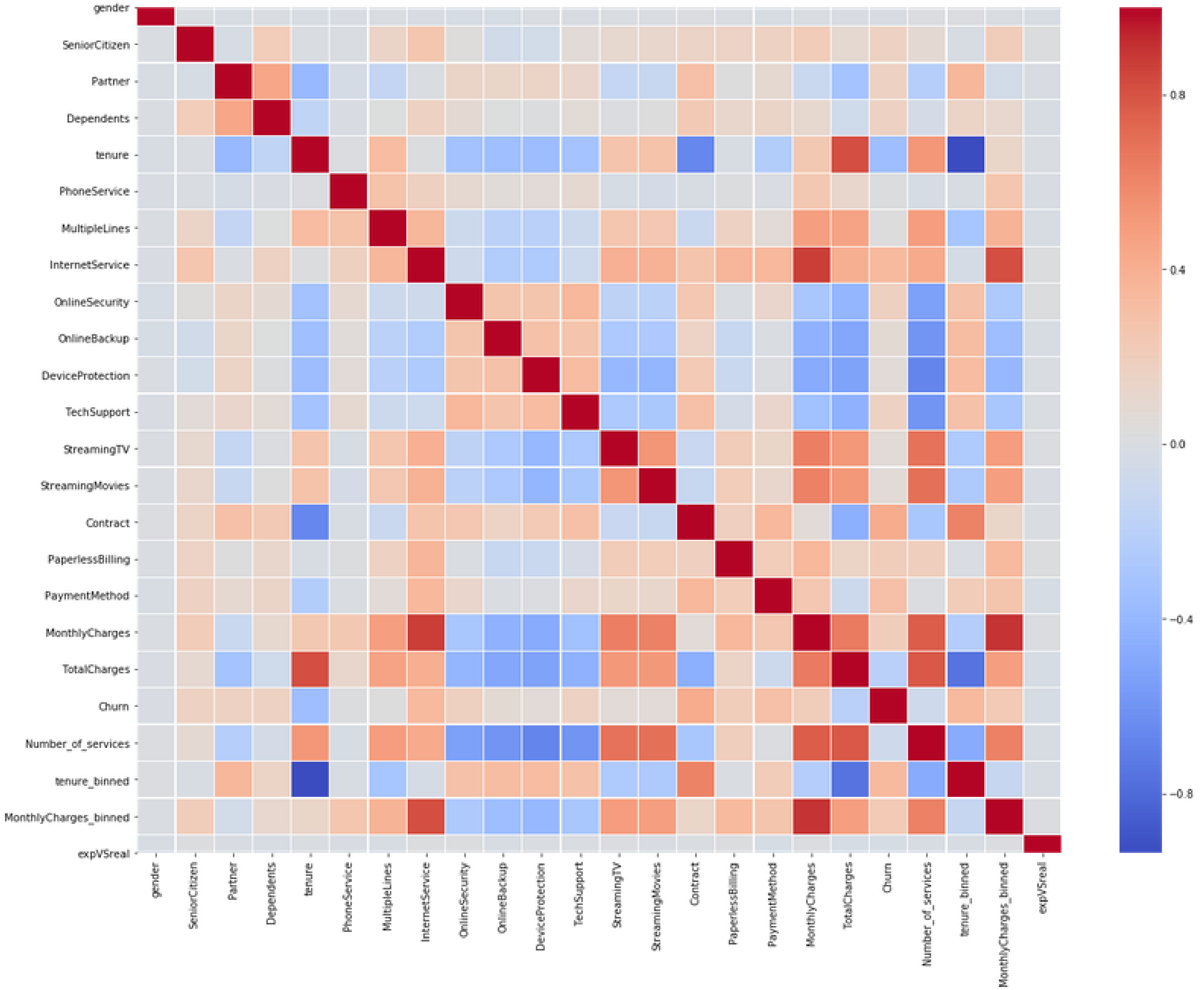
Correlation Matrix with One-Hot Encoding



Feature selection with One-Hot encoding:

- **Here we drop the variables which are perfectly negatively correlated with another variable:**
  - Gender Male
  - SeniorCitizen_No
  - Partner_Yes
  - Dependents_Yes
  - PhoneService_Yes
  - MultipleLines_No
  - OnlineSecurity_Yes
  - OnlineBackup_Yes
  - DeviceProtection_No
  - TechSupport_Yes
  - StreamingTV_Yes
  - StreamingMovies_Yes
  - PaperlessBilling_No

Correlation matrix with Mean Encoding



Feature selection with Mean encoding

- **Here we dropped because they are highly correlated with other variables:**
  - tenure_binned
  - Monthly_Charges

Section 5

Models

Selected Models

- Logistic Regression (l1 penalty)
- Logistic Regression (l2 penalty)
- Support Vector Classifier
- Random Forest
- Gradient Boosting
- XGBoost

We fitted each model twice: one for mean encoding and one for one-hot encoding

Pipeline

Train Test split

Upsampling

Grid Search for each model

Except for the XGBOOST

Train-Test Split and Upsampling

- **To solve unbalancedness of data upsampling is performed**

  This resamples randomly from the minority class in order to rebalance data and increase performances of models

- **Data split: 80% training and 20% test**

  Stratified shuffle split exploited to keep the ratio between Churning and not Churning clients

Grid Search

- **Performed grid search to fine tune hyperparameters of each model maximizing AUC score**

- **Fine tune the threshold of the hard classification to maximize F1 score**

Spotting Churn client is more relevant than spotting non Churn clients. By maximizing F1 we reach a good compromise between precision and recall

Section 6

Results

Choosing the best model

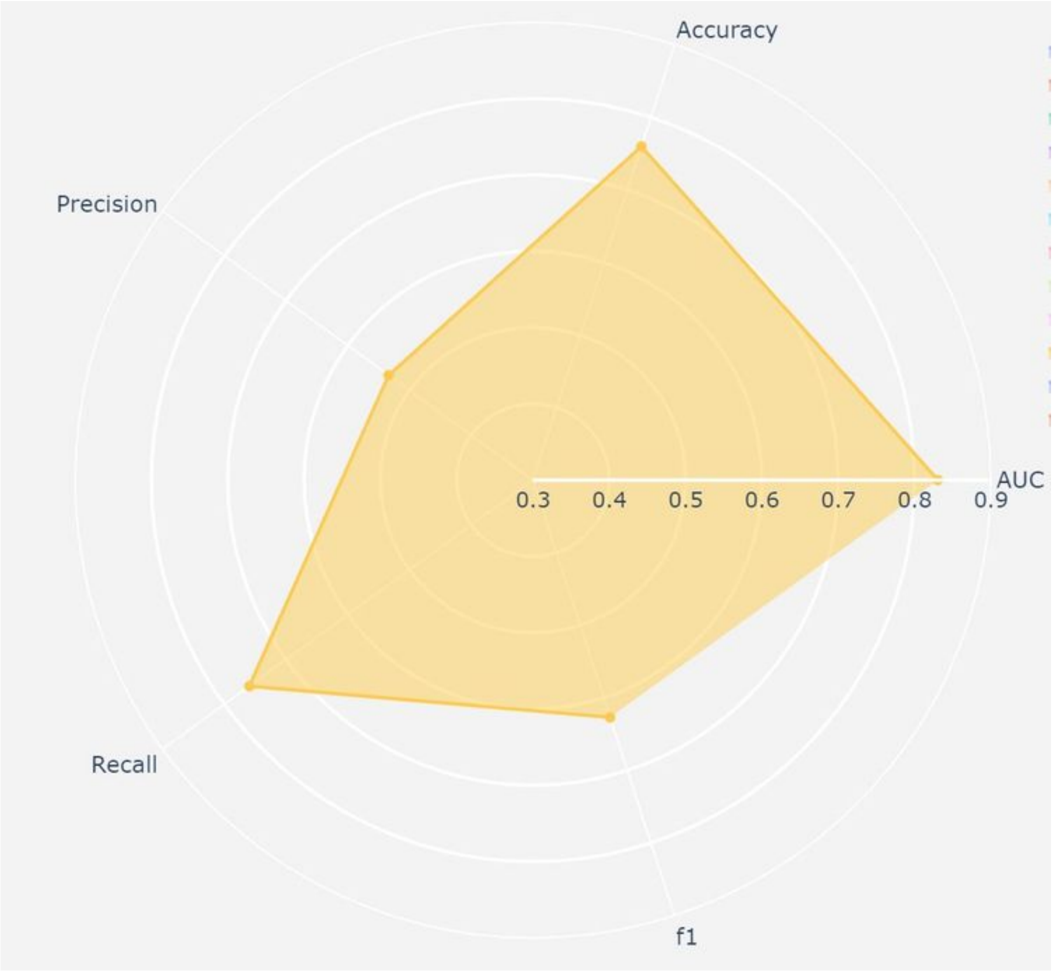
- **Plot that highlights the performance metrics of each model**

  See the notebook for the interactive plot

- **Highest recall reached by XGBoost model with One-hot encoding**

- **Highest F1 score reached by Random Forest with mean encoding**

The Selected best model



- **We choose the Random Forest given the high performance in terms of F1 score**
- **There is also a good trade off in terms of precision and recall**
- **Performances:**

  AUC 0.8306642

  Accuracy 0.7619047

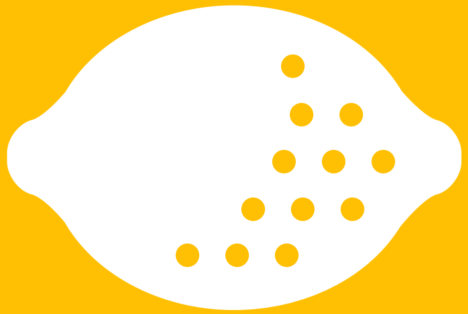
  Precision 0.5370018

  Recall 0.7566844

  F1 0.6281908

Section 7

Lime



Why Lime?

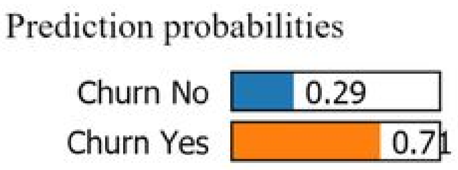
- **Allows to interpret black box models**

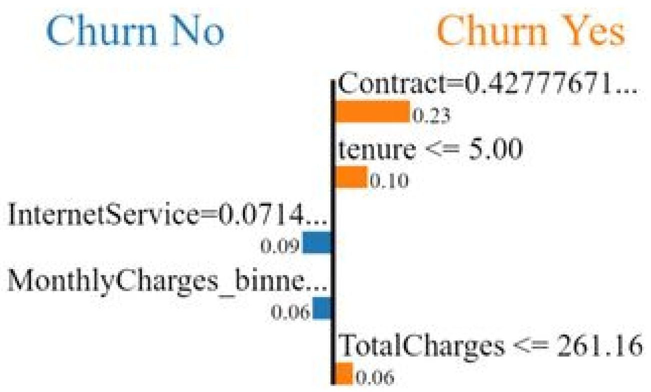
  Black Box models are difficult to interpret. Lime gives us a tool to interpret (at least locally) these kind of models

- **Provides Business a tool to understand how to deal with risky customers**

  Thanks to the output provided by this algorithm, the customer relationship management can understand what are the elements to leverage in order to retain risky clients

Illustrative **Example**





- ## The Customer selected is very likely to churn

  The features in orange in the second plot highlight which are the most impacting elements in determining the probability of churn

- ## Implications

  In this specific case for example, the company could offer the customer a new type of contract so that its churn probability is reduced

- ## Case by case tool of reaction

  Thanks to our model and this algorithm, the CRM is provided a tool to react properly by considering all the characteristics of each client

Section 8

Conclusions

Business conclusions

- **We found out that the best model is Random Forest with mean encoding and F1 score of 62.8%.**

- **We think that it can be improved if the company could add other type of data that identifies the customer engagement which are easy collectable , for example:**
  - The number of claims that each client does
  - The number of times a client contacted the call center
  - A survey on which each client express his opinion about the services given by the company

- **With Lime we provide a useful tool in order to deal better with potential churning clients. Identifying and tackling the critical variables is easier for the management of the company thanks to the model we created.**