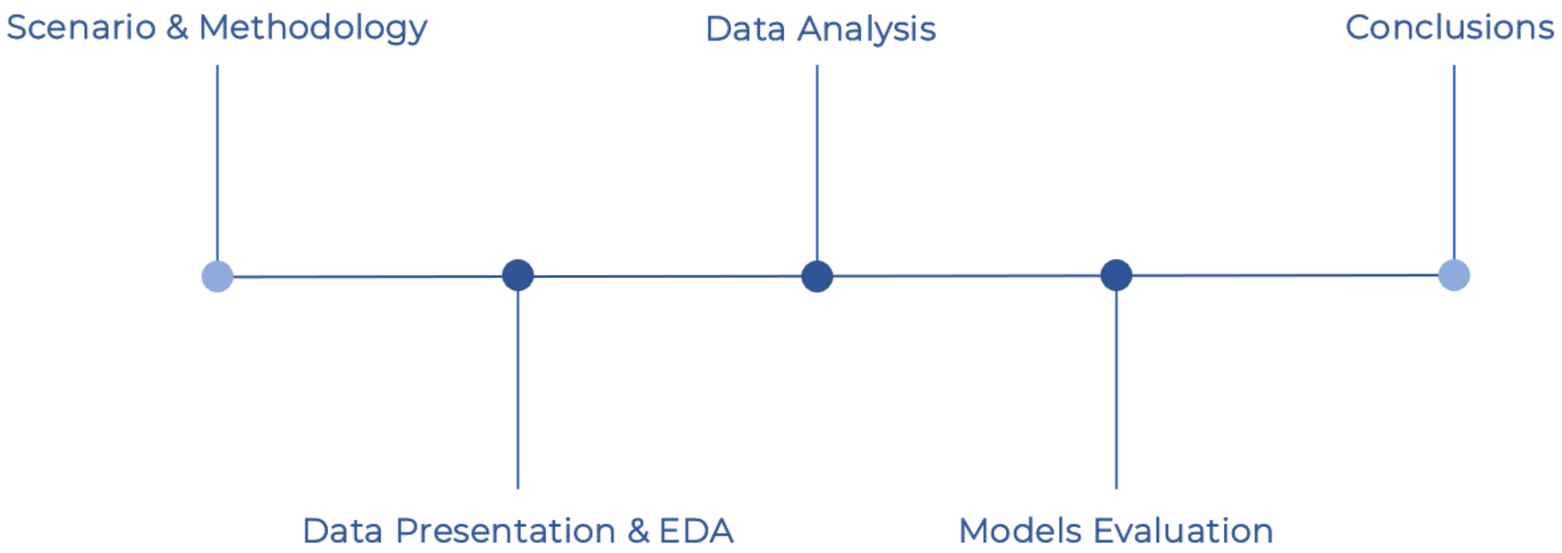
MACHINE LEARNING FOR CHURN PREDICTION

An analysis over Telco customers

Data Mining Project - AA 2019-2020 UCSC

Massimo Checchin, Andrea Corvi, Edoardo Malerba, Alessandro Noseda

Agenda


Scenario & Methodology
Data Presentation & EDA
Data Analysis
Models Evaluation
Conclusions

Project summary

- Stakeholder:
  Telecommunication company

- Objective:
  Building the best machine learning model to predict customer churn

- Methodology:
  Compare six different binary classification models in order to predict the outcome for each client.
  - **Logistic Regression**
  - **Decision Tree**
  - **Random Forest**
  - **Gradient Boosting**
  - **SVM**
  - **Ensemble**

- Evaluation:
  Model tuning: AUC
  Final model selection: F1 Score

Objective Explaination

- What is churn?

  Customer churn occurs when an existing customer, user, player, subscriber or any kind of return client stops doing business or ends the relationship with a company. It is a key business metrics because often the cost of retaining an existing customer is far less than acquiring a new one.

- Why is a model for predicting churn useful?
  - Because if a company is able to predict if a customer will leave, then it can intervene to retain it.
  - Besides building the best predictive model, our objective is also to understand what are the key factors that lead a client to this choice.

Provided Data

- Binary target variable: "Churn"

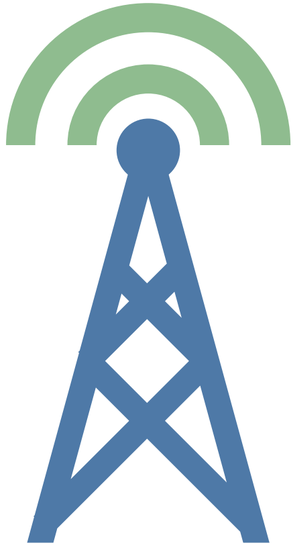
- Dataset Dimension: 7043 rows, 21 variables:

Services variables:

- Tenure
- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod
- MonthlyCharges
- TotalCharges

Customer information variables:

- CustomerID
- Gender
- SeniorCitizen
- Partner
- Dependents



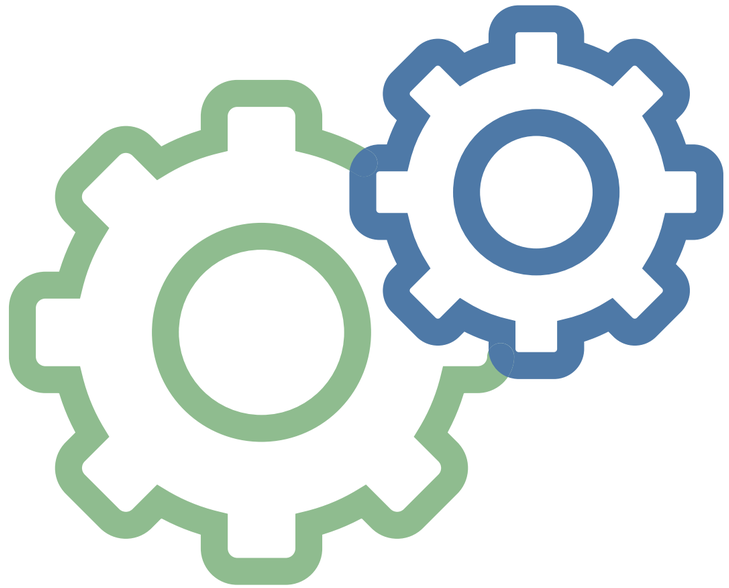
Data Preprocessing

- Imputation of missing data:
  Eleven rows had no data about "Total Charges" while having "Tenure" = 0.
  Since Tenure is the number of billed month, missing values are replaced with 0.

- Feature Engineering:
  - **Merge "PhoneService" and "MultipleLines" into a new variable "PhoneLines" since combined they give the same information.**
  - **Create a new column "TotalServices" as the sum of all the active services (such as: OnlineSecurity, StreamingTv, etc.) for each client.**

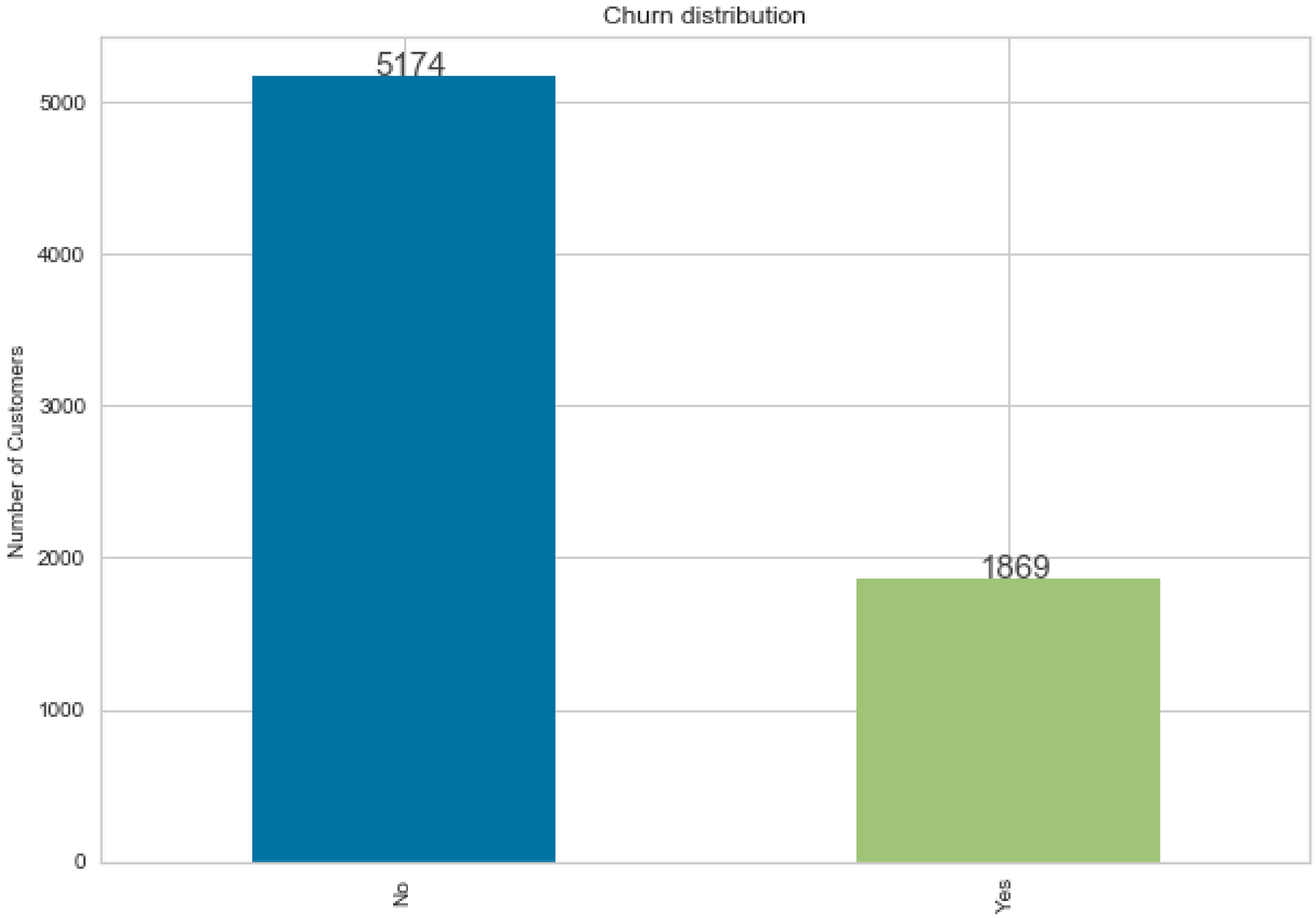
- Drop "CustomerId" column:
  Since it is not useful for the analysis

- Creation of dummies for categorical variables.

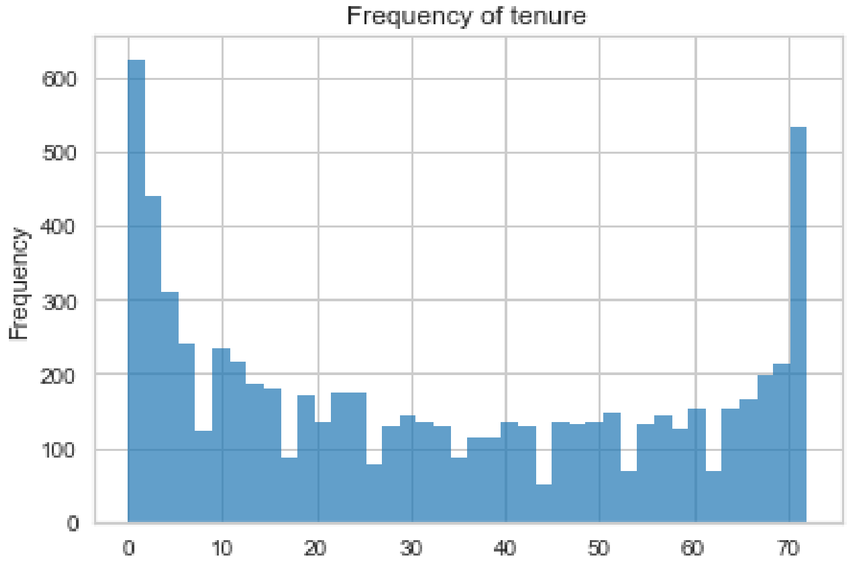


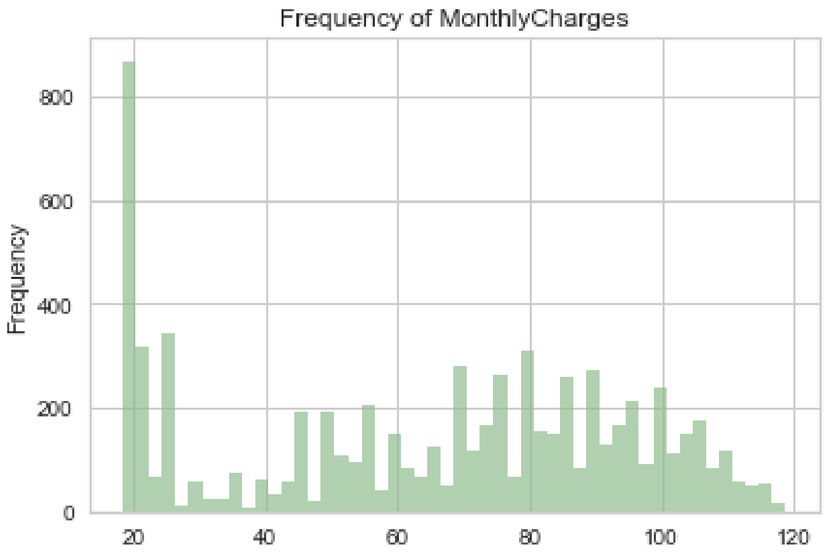
EDA - Target

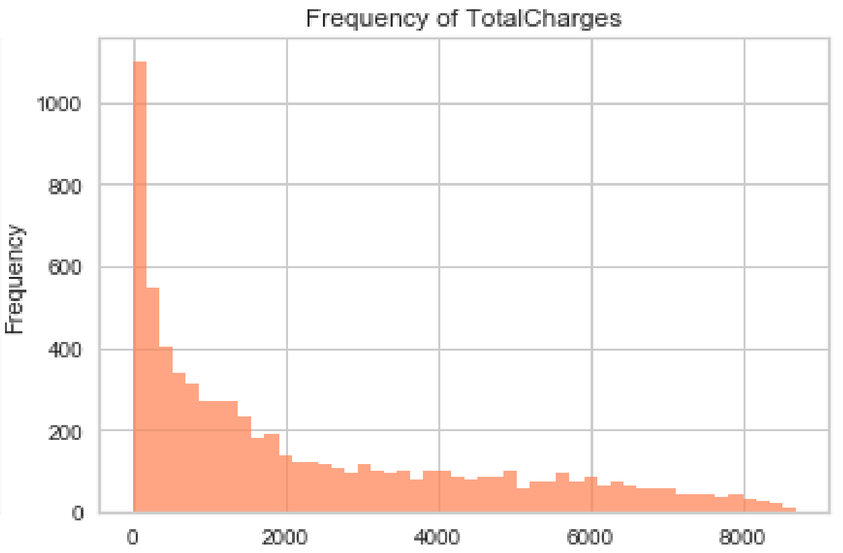
About 26% of the clients base has left the company.
The data is not evenly balanced: we may have to change the probability threshold during the classification.

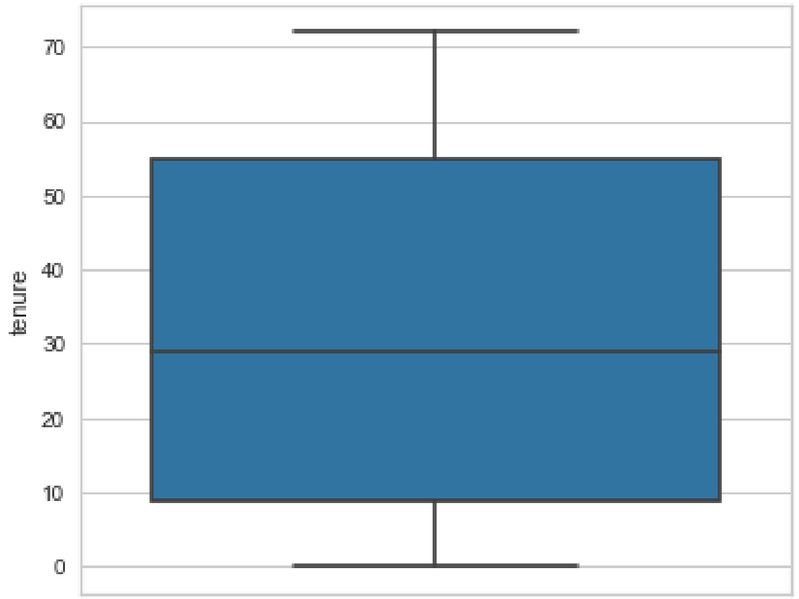


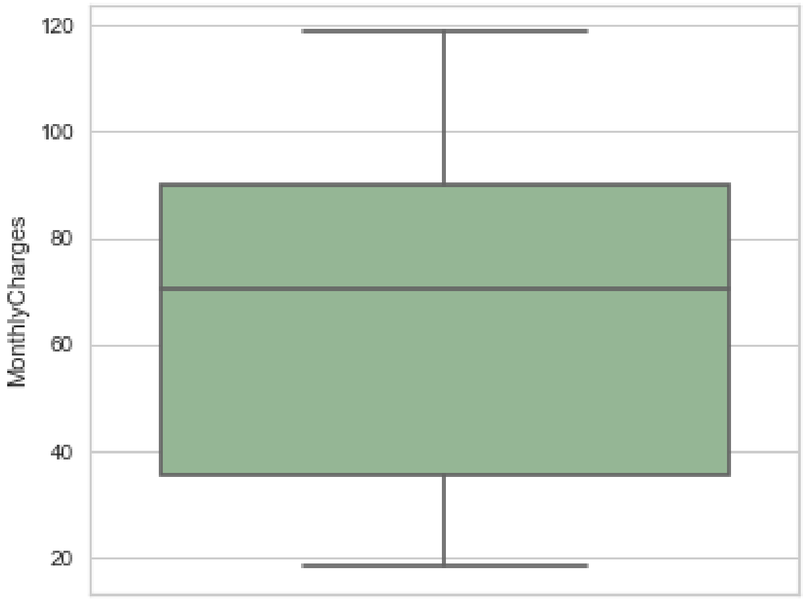
EDA - Numerical variables

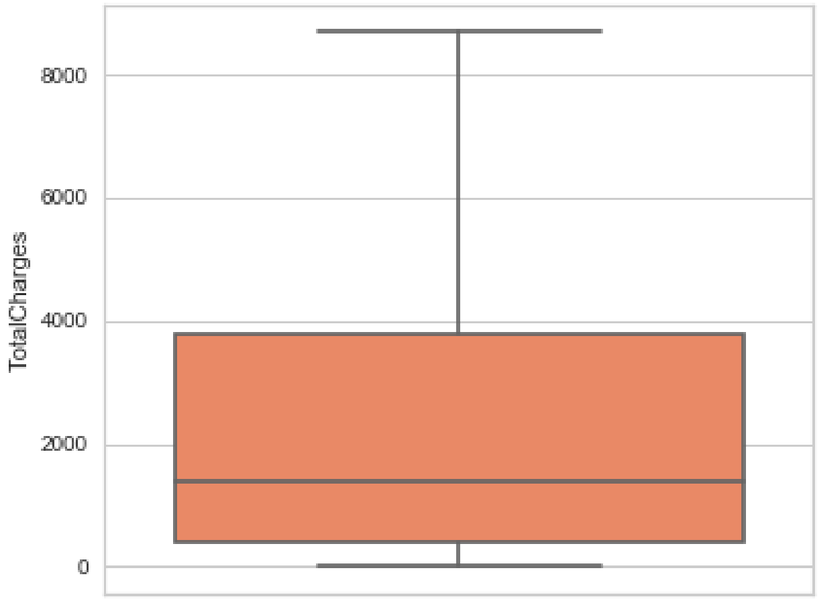




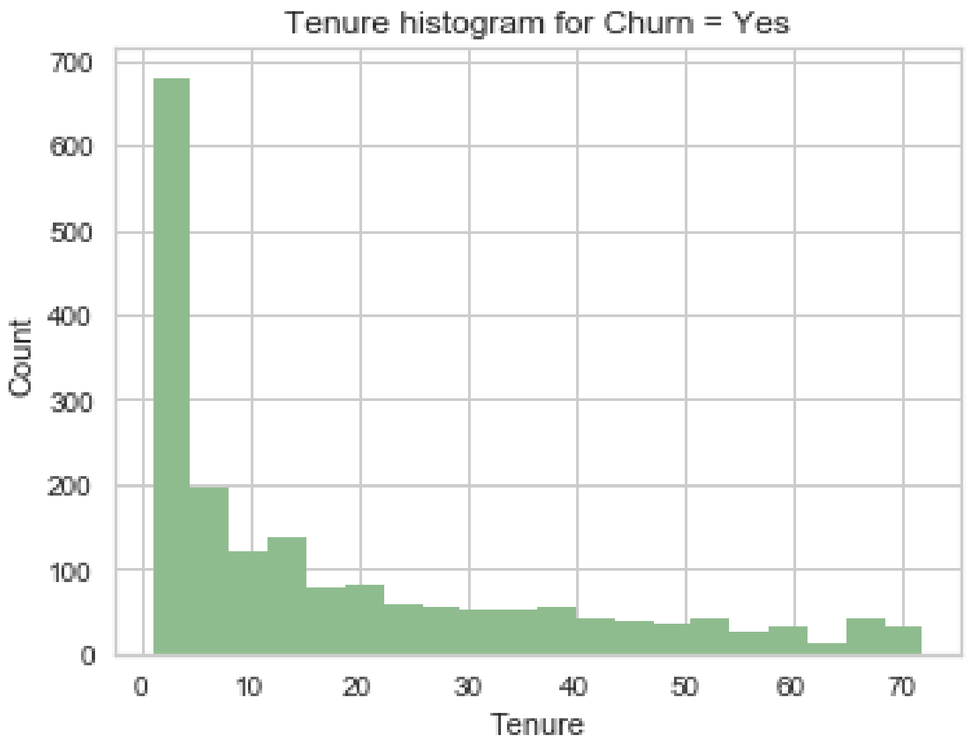


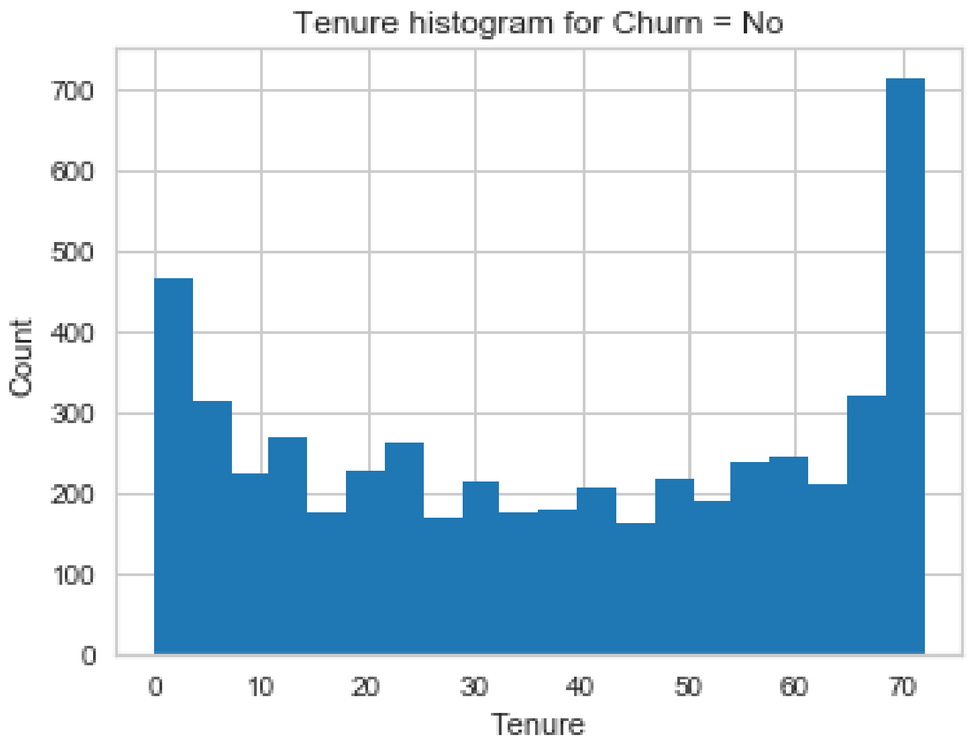






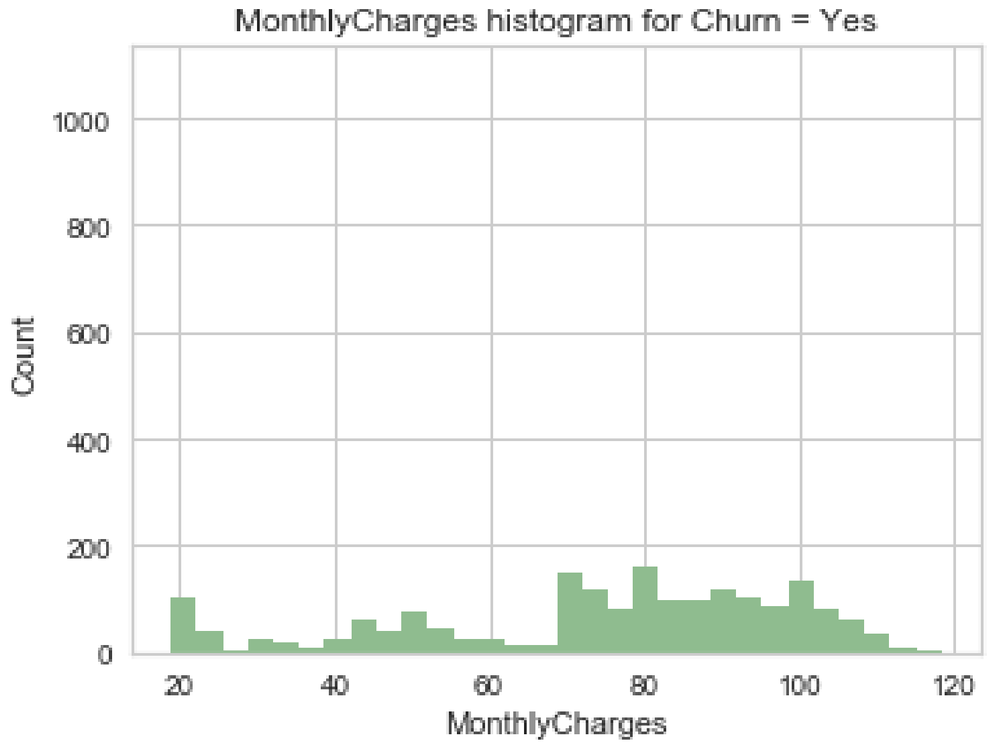
EDA - Tenure

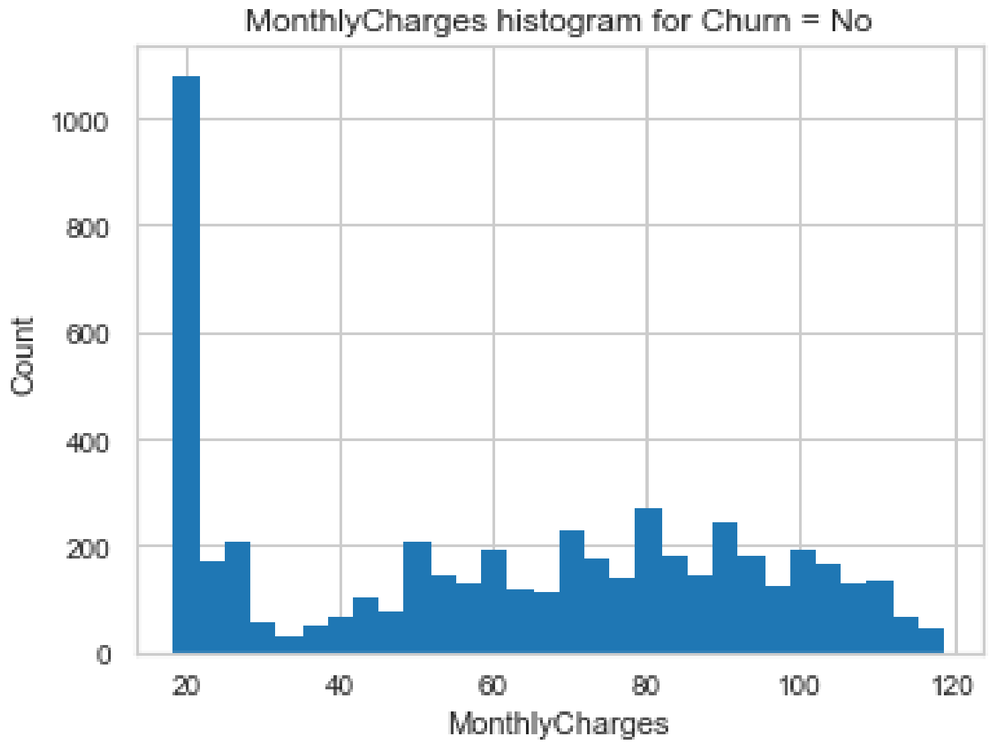




From the plot it's possible to see that the distribution of Tenure for clients that leave is skewed. It implies that the more one person is a customer of the company, the more he/she will be likely to stay within the current provider.

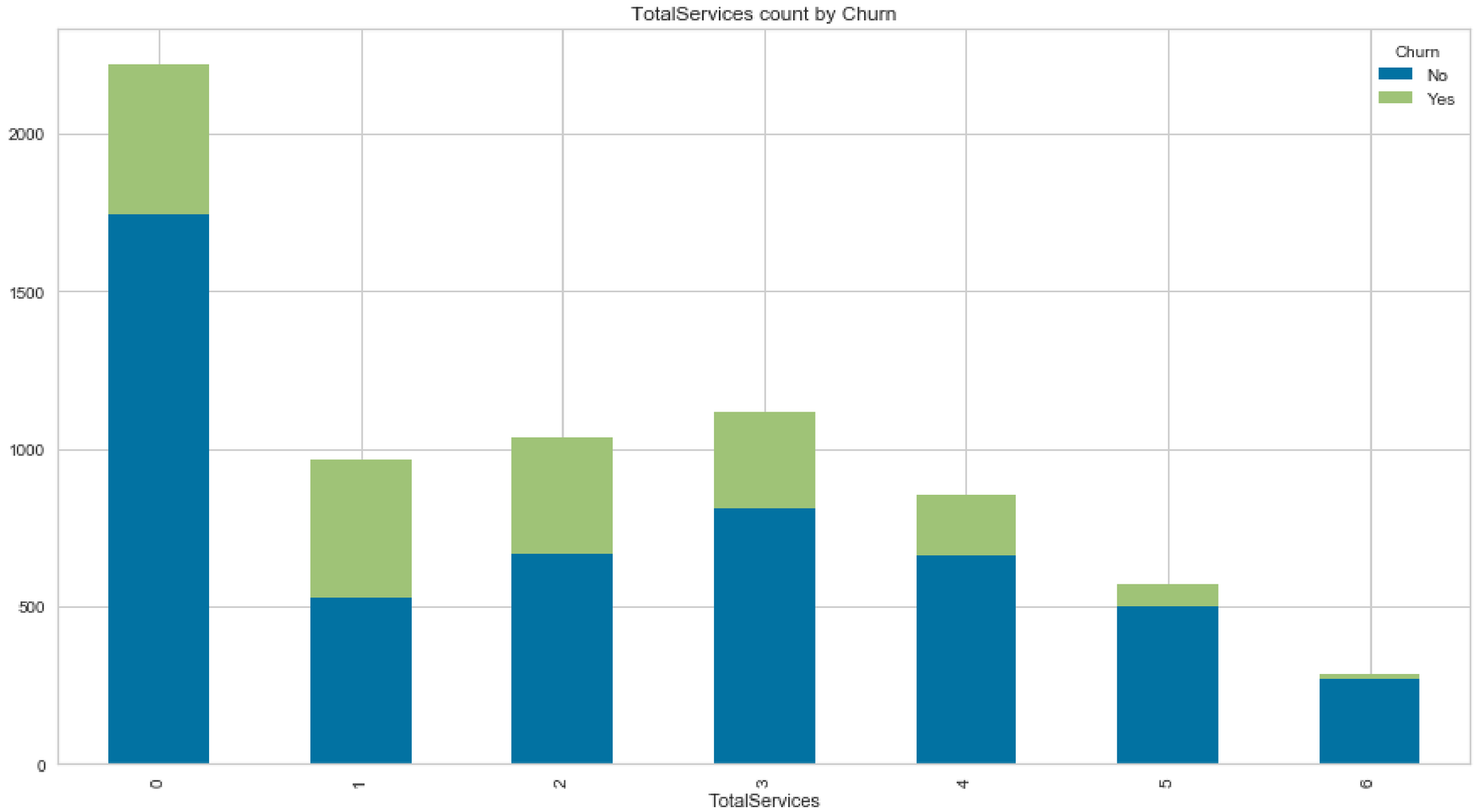
EDA - Monthly Charges





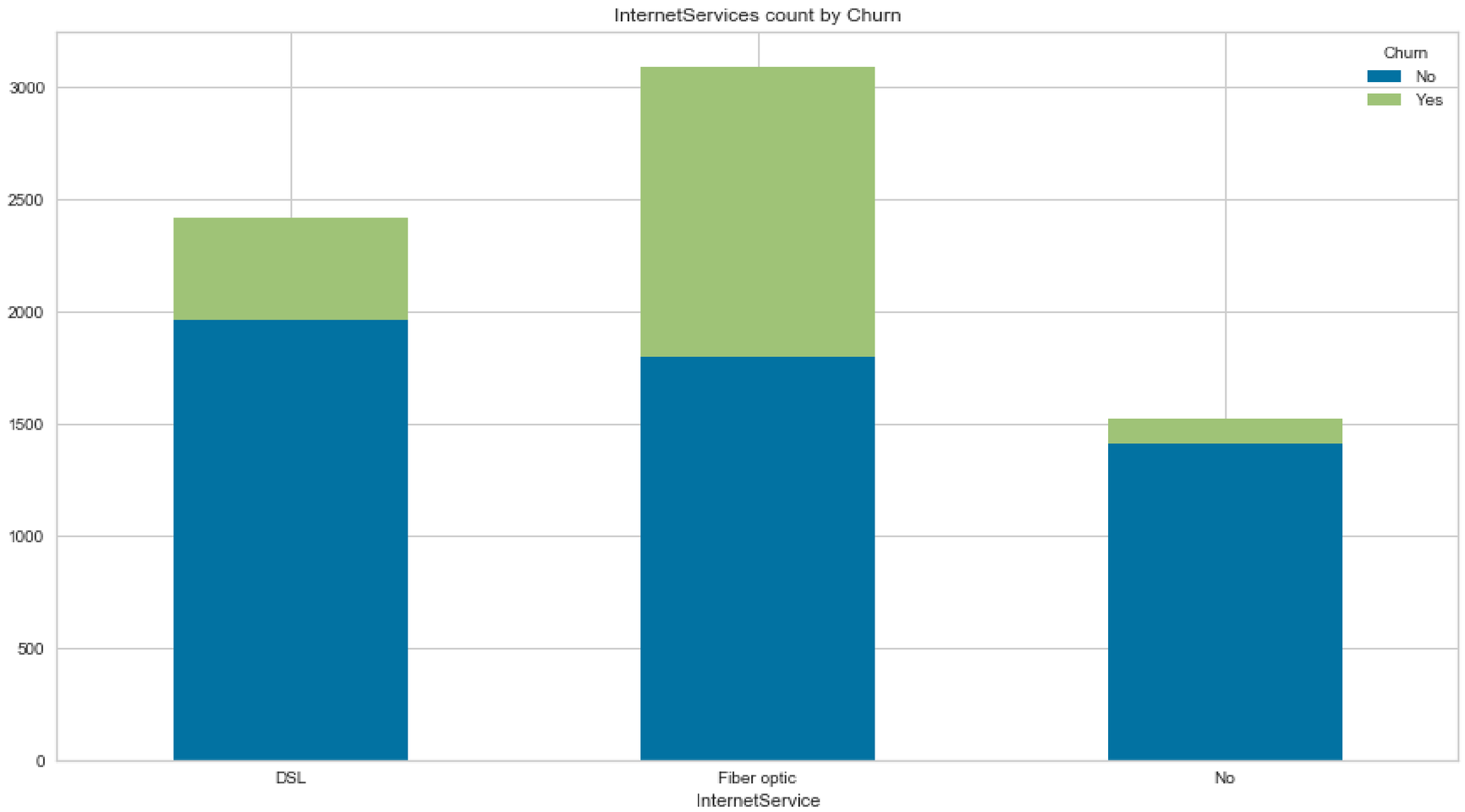
From the plot it's possible to see that if the monthly charge is low (e.g. less than 30), the number of customers that leave is way smaller than the number of clients that remain with the company. We expect so that "MonthlyCharges" will be a strong predictor during the analysis.

EDA - Total Service



Given the distribution shown on the plot, we suppose that clients with many services are more likely to stay.

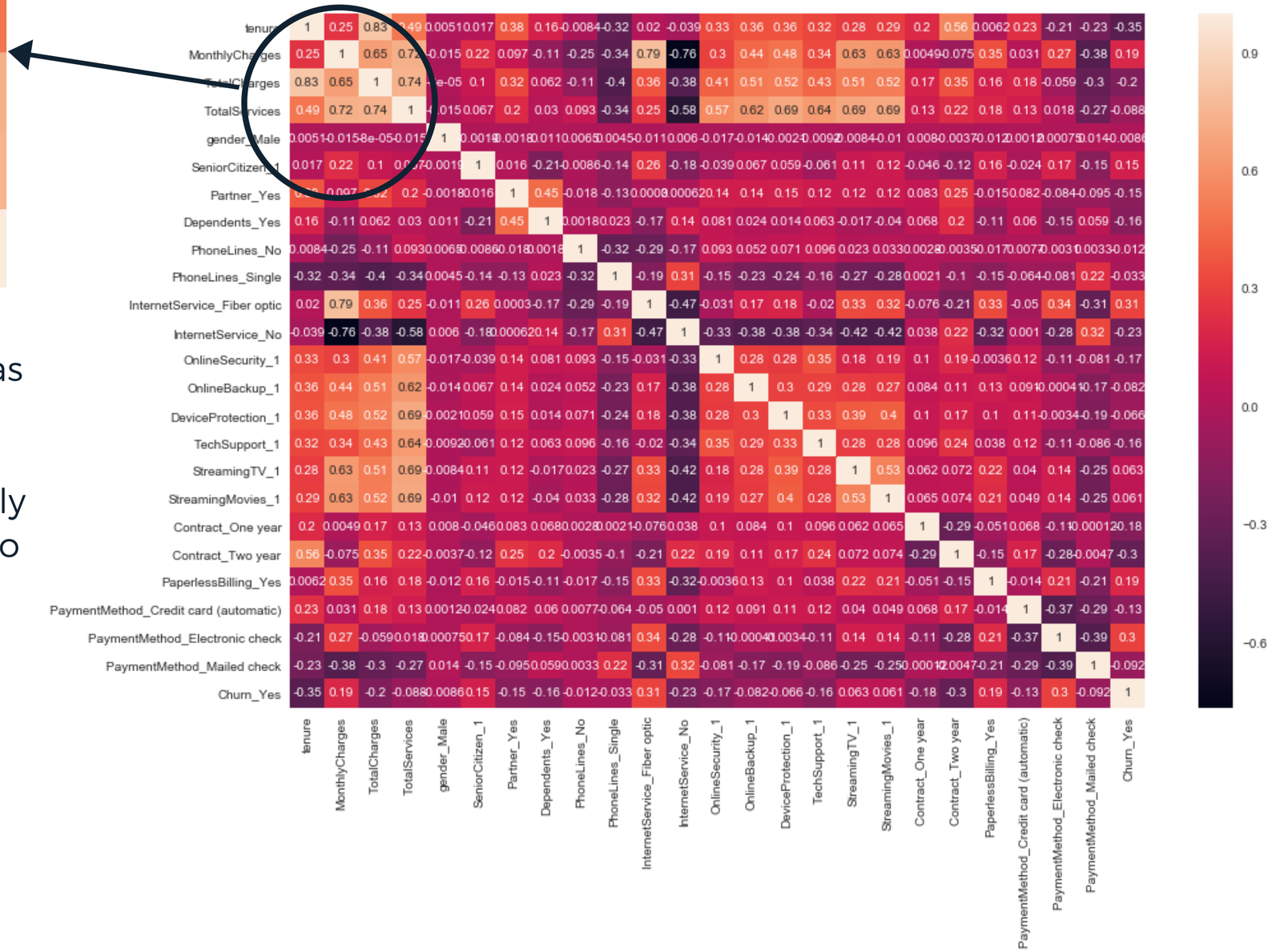
EDA - InternetServices



One interesting insight that we can understand from the plot is that if a client doesn't have an internet connection, then he's more likely to stay. Our guess is that this kind of client could be elders and so not prone to switch providers.

EDA - Correlation Matrix

tenure	1	0.25	0.83	0.49
MonthlyCharges	0.25	1	0.65	0.72
TotalCharges	0.83	0.65	1	0.74
TotalServices	0.49	0.72	0.74	1



Since "TotalCharges" is almost equal as the product of tenure and monthly charges (plus some costs such as: activation/installation costs), it is highly correlated with those two variables. So we decided to drop it, alongside with the variables related to the internet services (since we already have total services) for the fitting of the logistic regression,

Analysis

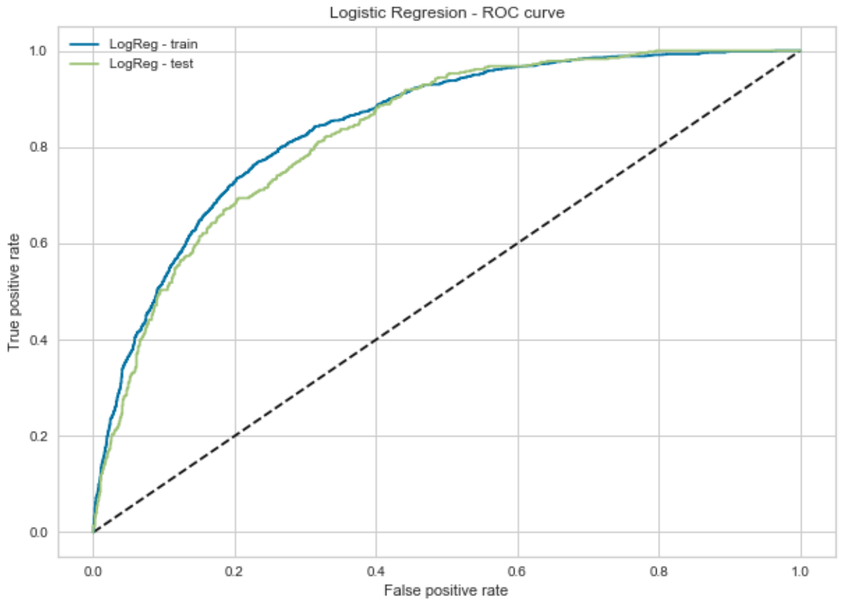
- Regarding the fitting and evaluation of our models, we decided to split the dataset into:
  - Train set = 75% of the data.
  - Test set = 25%.

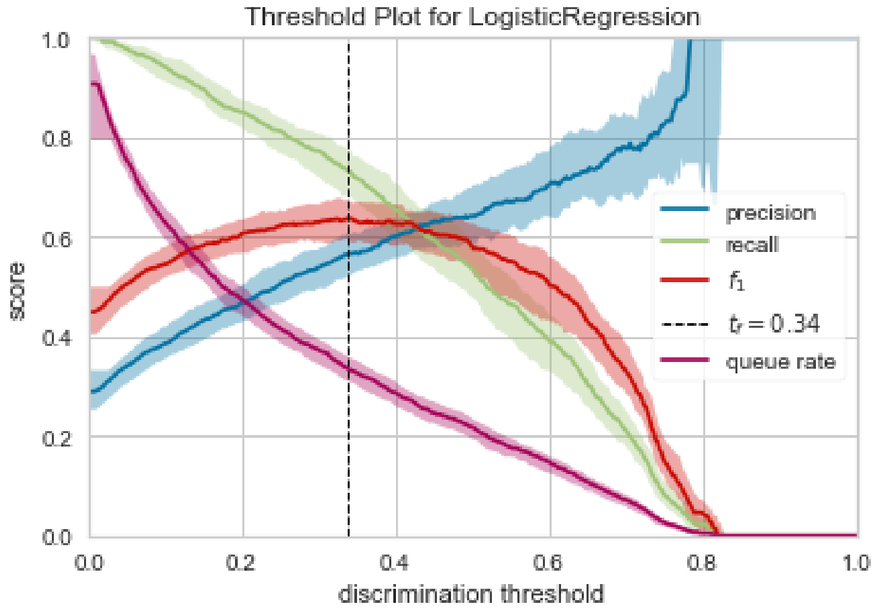
- During the tuning of the parameters on the train data we used 5-fold cross validation.

- Models:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosting
  - SVM
  - Ensemble

- The classification threshold is chosen by maximizing the F1 Score

Logistic Regression



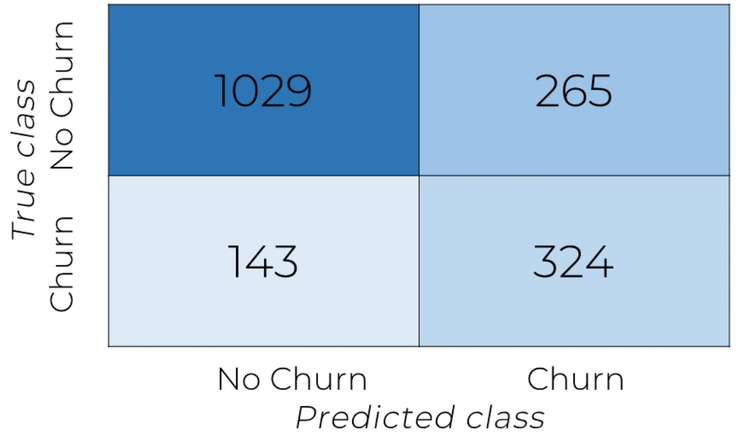


Area under the curve:

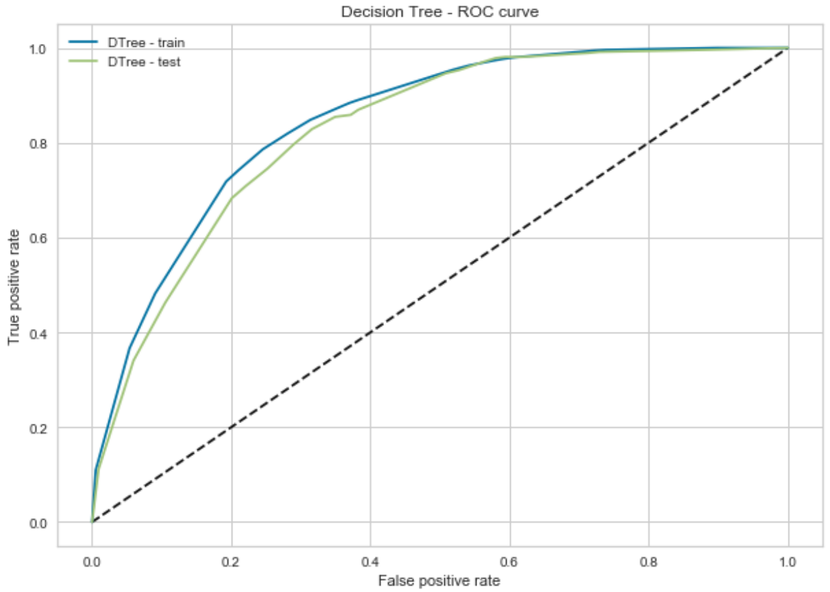
- 0.8432 Train AUC
- 0.8302 Test AUC

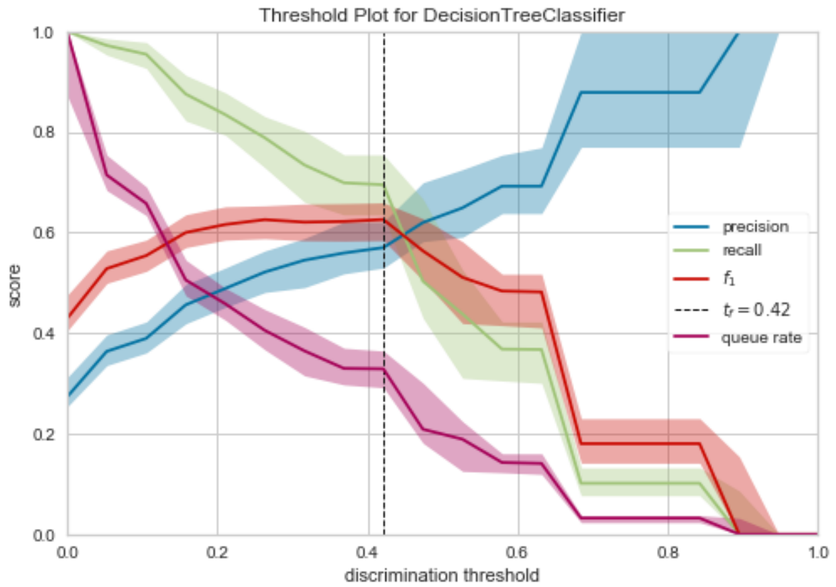
Metrics (threshold = 0.34):

- 0.5500 Precision
- 0.6937 Recall
- 0.7683 Accuracy
- 0.6136 F1 Score



Decision Trees



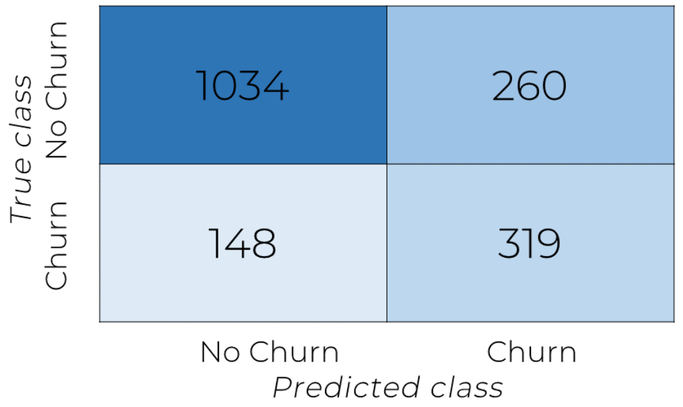


Area under the curve:

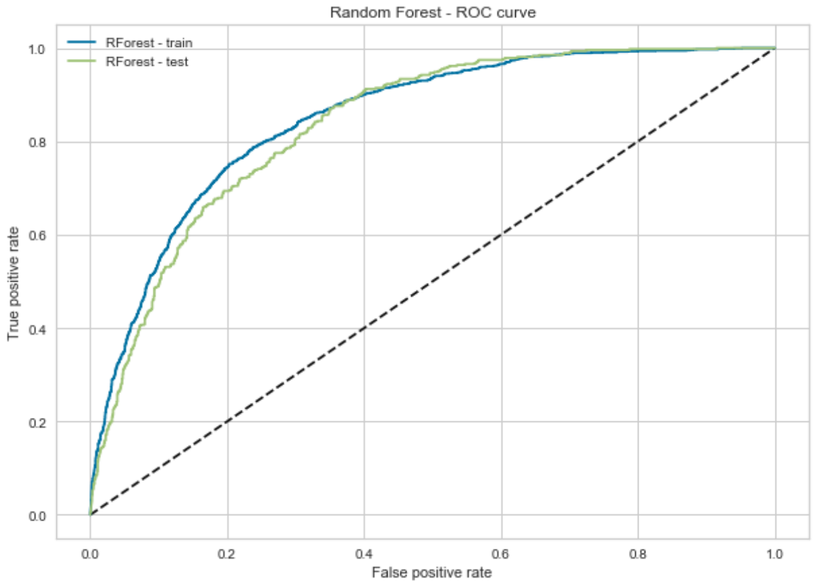
- 0.8462 Train AUC
- 0.8281 Test AUC

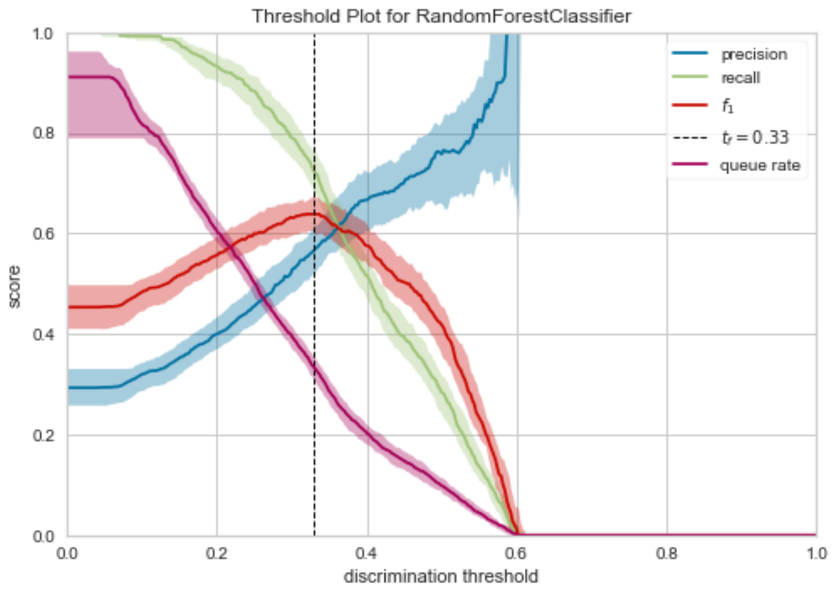
Metrics (threshold = 0.42):

- 0.5509 Precision
- 0.6830 Recall
- 0.7683 Accuracy
- 0.6099 F1 Score



Random Forest



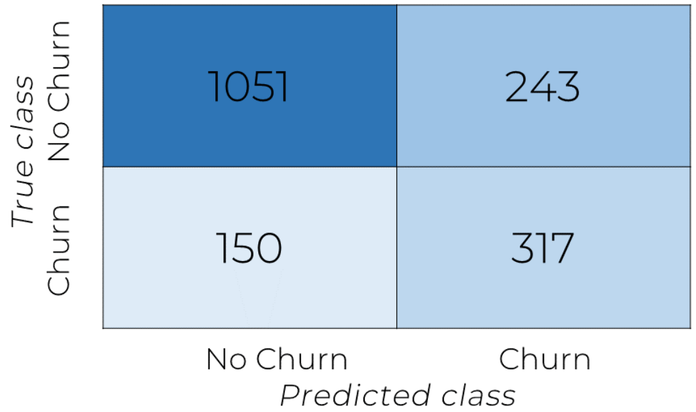


Area under the curve:

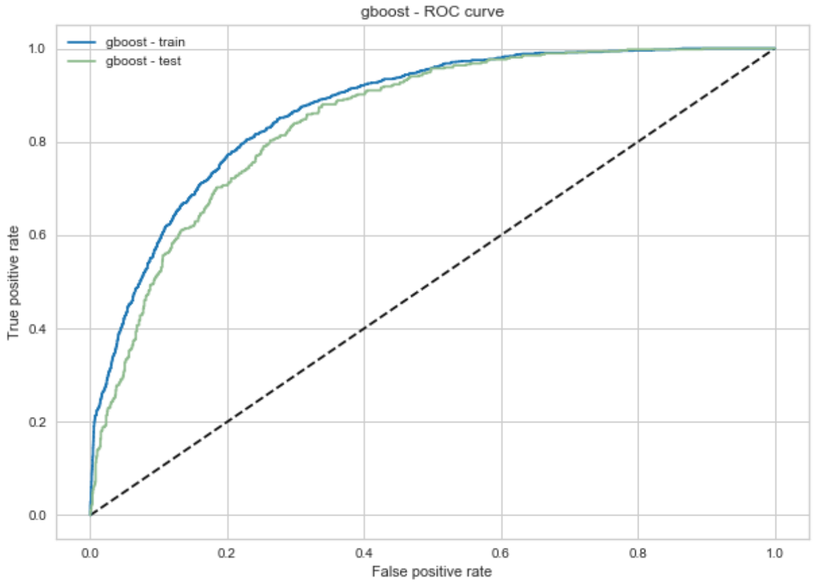
- 0.8493 Train AUC
- 0.8374 Test AUC

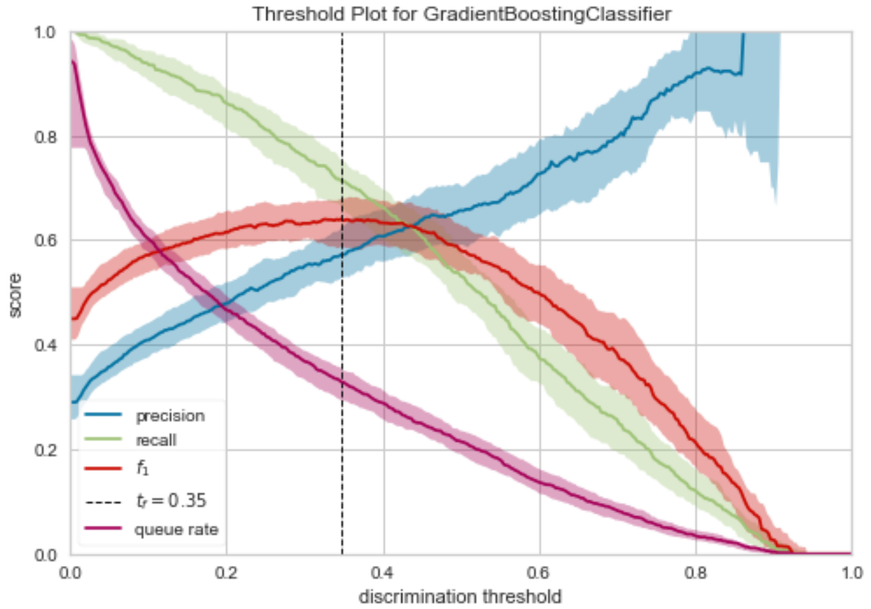
Metrics (threshold = 0.33):

- 0.5660 Precision
- 0.6788 Recall
- 0.7768 Accuracy
- 0.6173 F1 Score



Gradient Boosting (Scikit-learn)



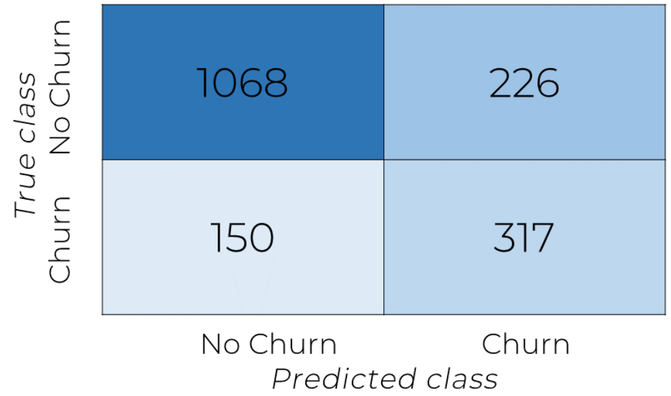


Area under the curve:

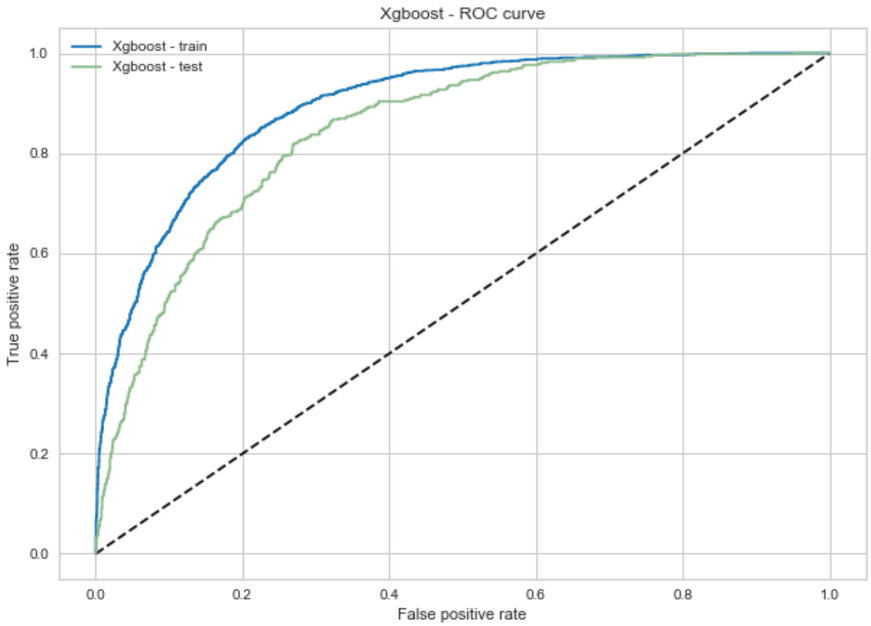
- 0.8690 Train AUC
- 0.8462 Test AUC

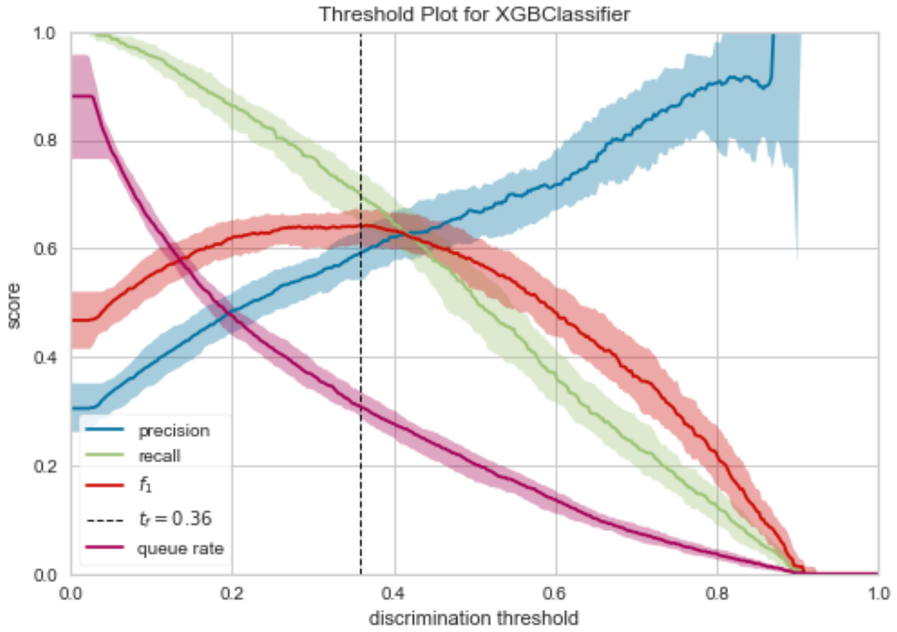
Metric (threshold = 0.35)

- 0.5837 Precision
- 0.6788 Recall
- 0.7864 Accuracy
- 0.6277 F1 Score



Xgboost



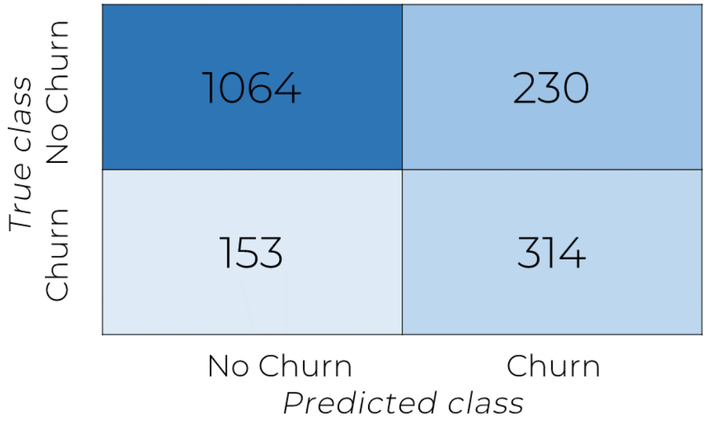


Area under the curve:

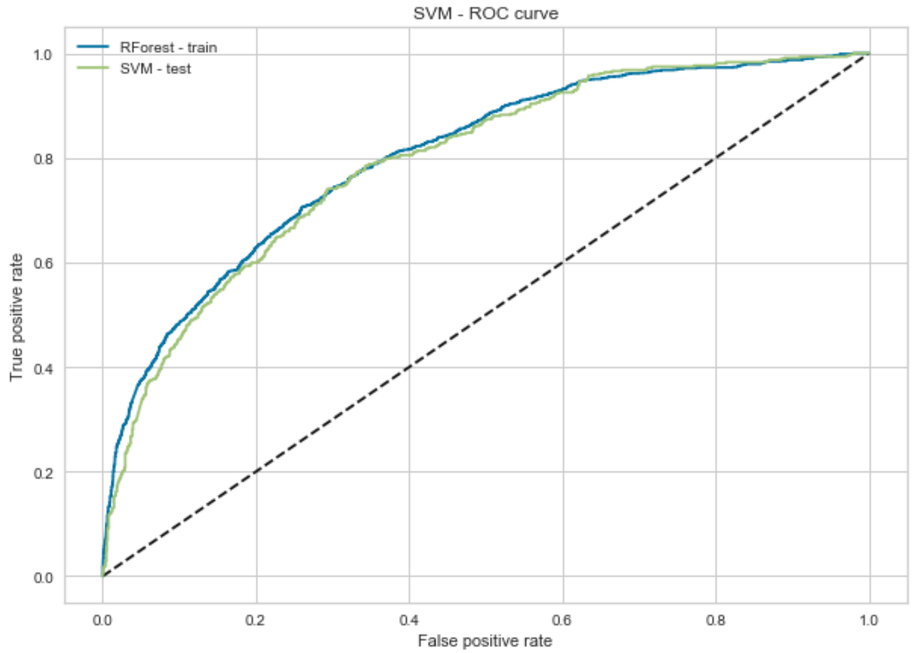
- 0.9002 Train AUC
- 0.8448 Test AUC

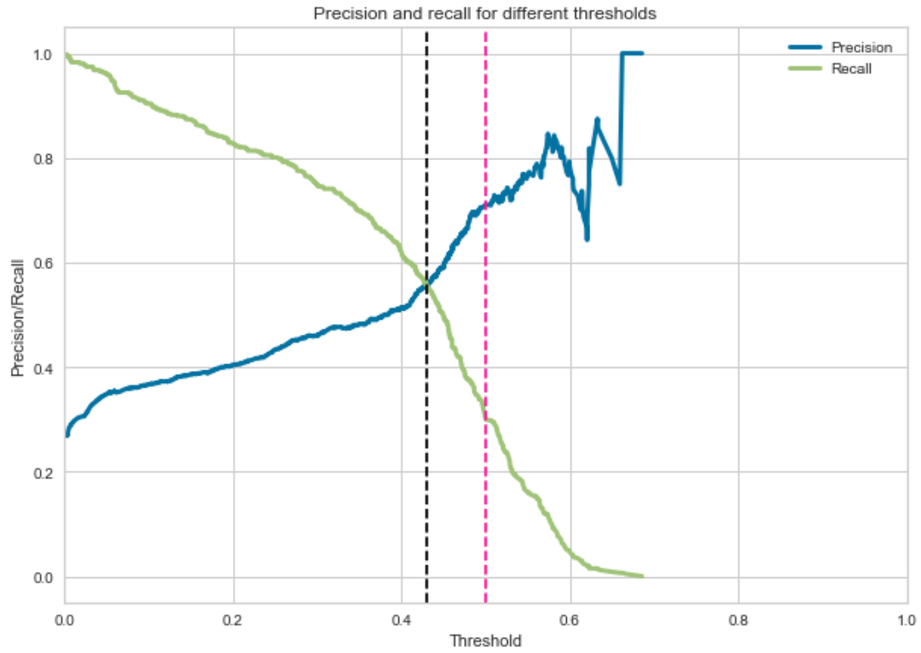
Metric (threshold = 0.36)

- 0.5772 Precision
- 0.6723 Recall
- 0.7825 Accuracy
- 0.6211 F1 Score



Support Vector Machines



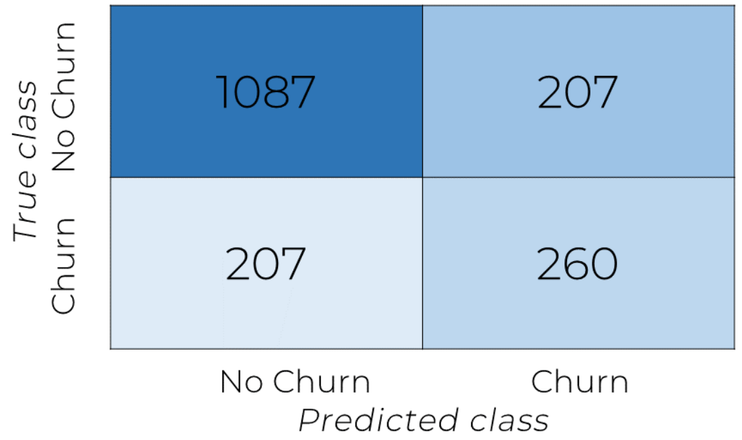


Area under the curve:

- 0.8014 Train AUC
- 0.7924 Test AUC

Metric (threshold = 0.43)

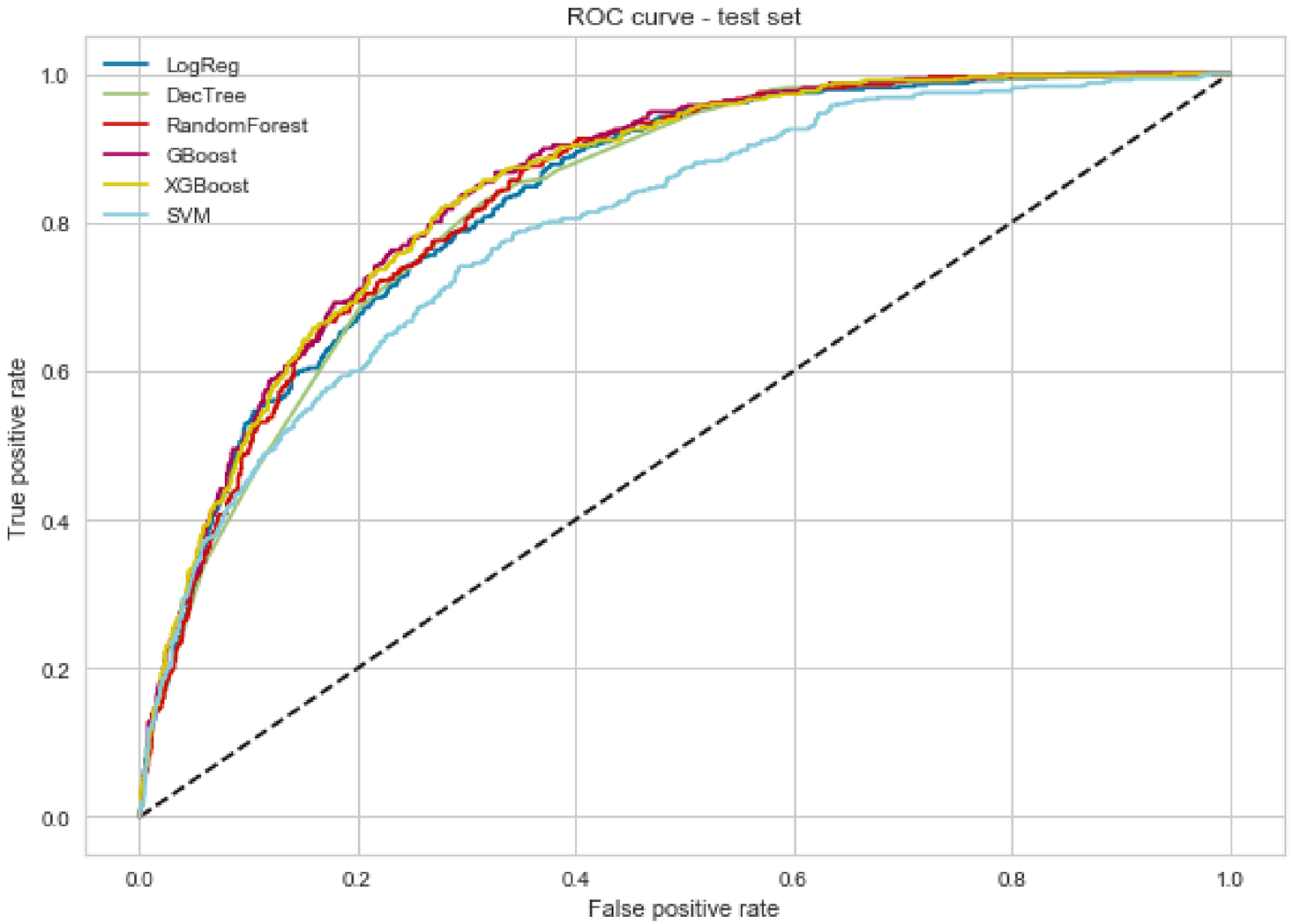
- 0.5567 Precision
- 0.5567 Recall
- 0.7649 Accuracy
- 0.5567 F1 Score



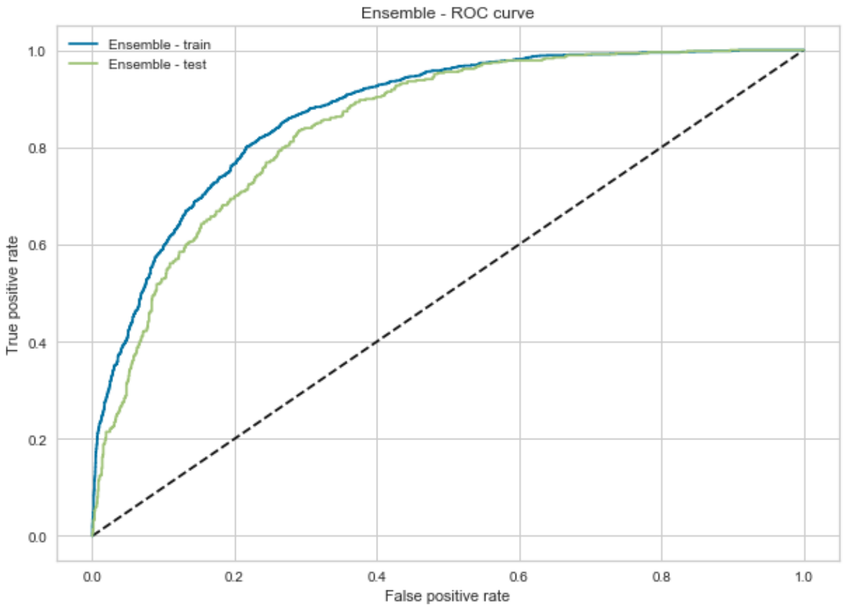
Models Comparison

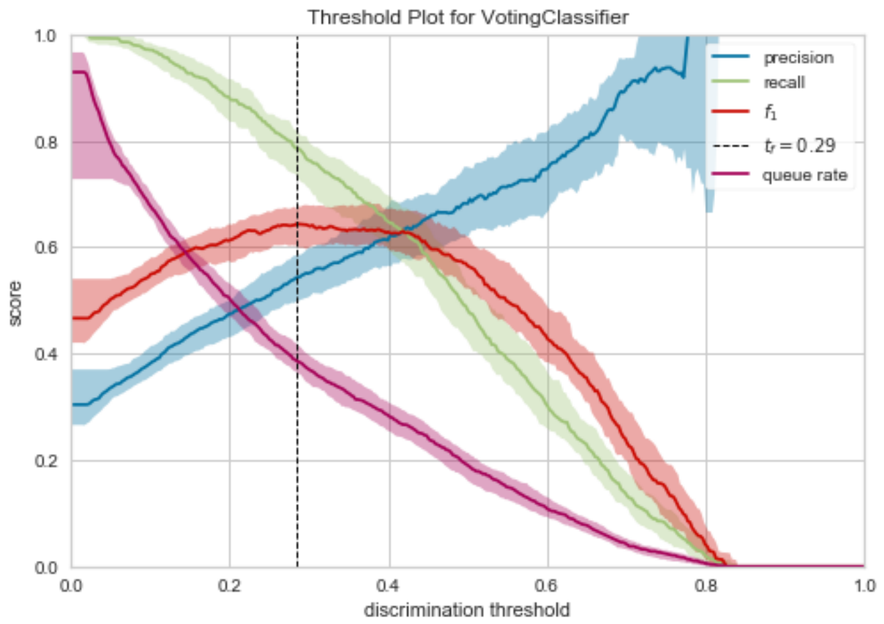
Also by looking at the plot of all the Roc curves for the test set, it is clear that one model, Support Vector Machines, is underperforming.

Because of this lack of performance and computational complexity we decided not to include the SVM model into the ensemble one.



Ensemble Model



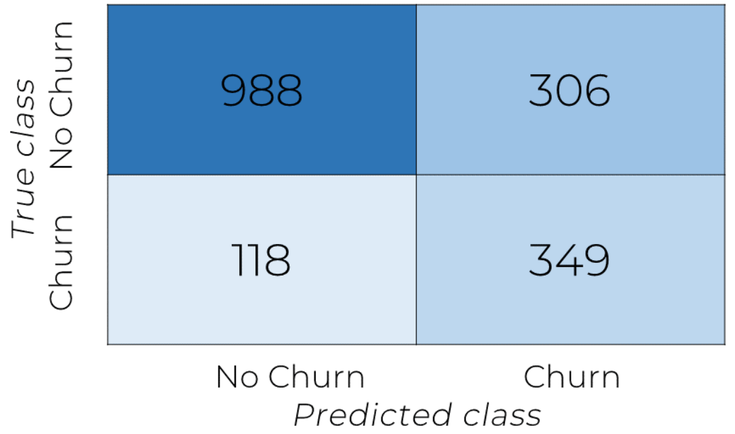


Area under the curve:

- 0.8719 Train AUC
- 0.8454 Test AUC

Metric (threshold = 0.29)

- 0.5328 Precision
- 0.7473 Recall
- 0.7592 Accuracy
- 0.6221 F1 Score



Final Evaluation

	auc_test	auc_train	f1_test	f1_train	precision_test	precision_train	recall_test	recall_train	accuracy_test	accuracy_train
lr	0.833082	0.847279	0.605364	0.636220	0.547660	0.569656	0.676660	0.720399	0.766042	0.781333
dt	0.828135	0.846242	0.609943	0.637947	0.550950	0.573789	0.683084	0.718260	0.768313	0.783605
rf	0.837413	0.849347	0.617332	0.647096	0.566071	0.580408	0.678801	0.731098	0.776831	0.788338
gb	0.847073	0.870571	0.627723	0.661717	0.583794	0.604363	0.678801	0.731098	0.786485	0.801590
xgb	0.844856	0.900244	0.621167	0.703826	0.577206	0.654601	0.672377	0.761056	0.782510	0.829989
svm	0.792491	0.801407	0.556745	0.568287	0.556745	0.565279	0.556745	0.571327	0.764906	0.769595
ensemble	0.845402	0.871972	0.622103	0.664127	0.532824	0.563682	0.747323	0.808131	0.759228	0.783037

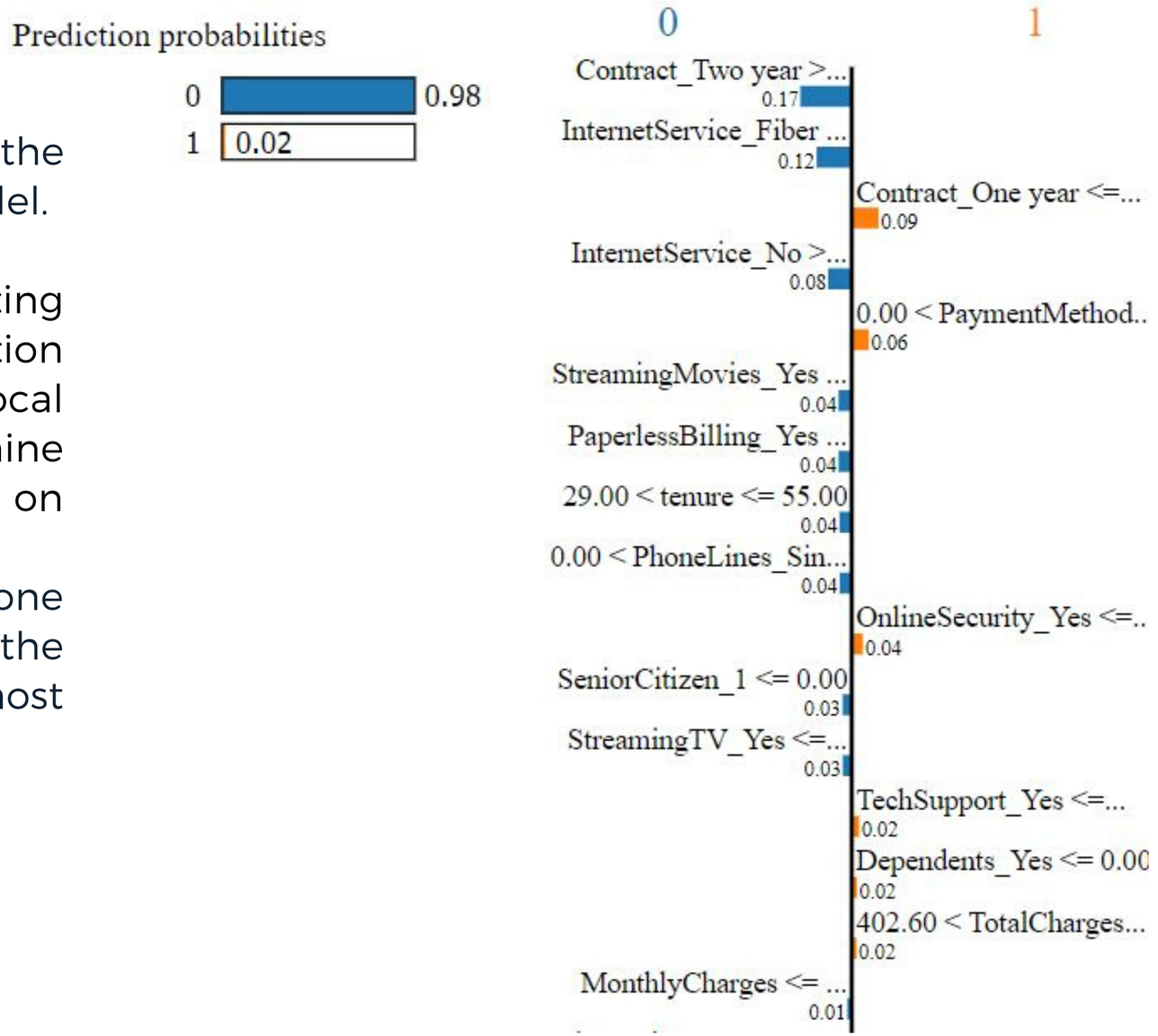
By looking at the chosen metrics, such as F1 Test and AUC Test, we can conclude that Gradient Boosting is the best model for predicting the churn of a current customer.
The ensemble model reaches an F1 Score close to the Gradient Boosting algorithm but given the fact that the ensemble is built on top of the GB Model, we prefer to keep a simpler model.

Although boosting is an effective technique for prediction, it has a downside which is the lack of interpretability. To our aid come two libraries of so called "explainers": "Lime" and "Shap".

Lime Explainer

Lime is a technique built in order to explain the prediction made by any machine learning model.

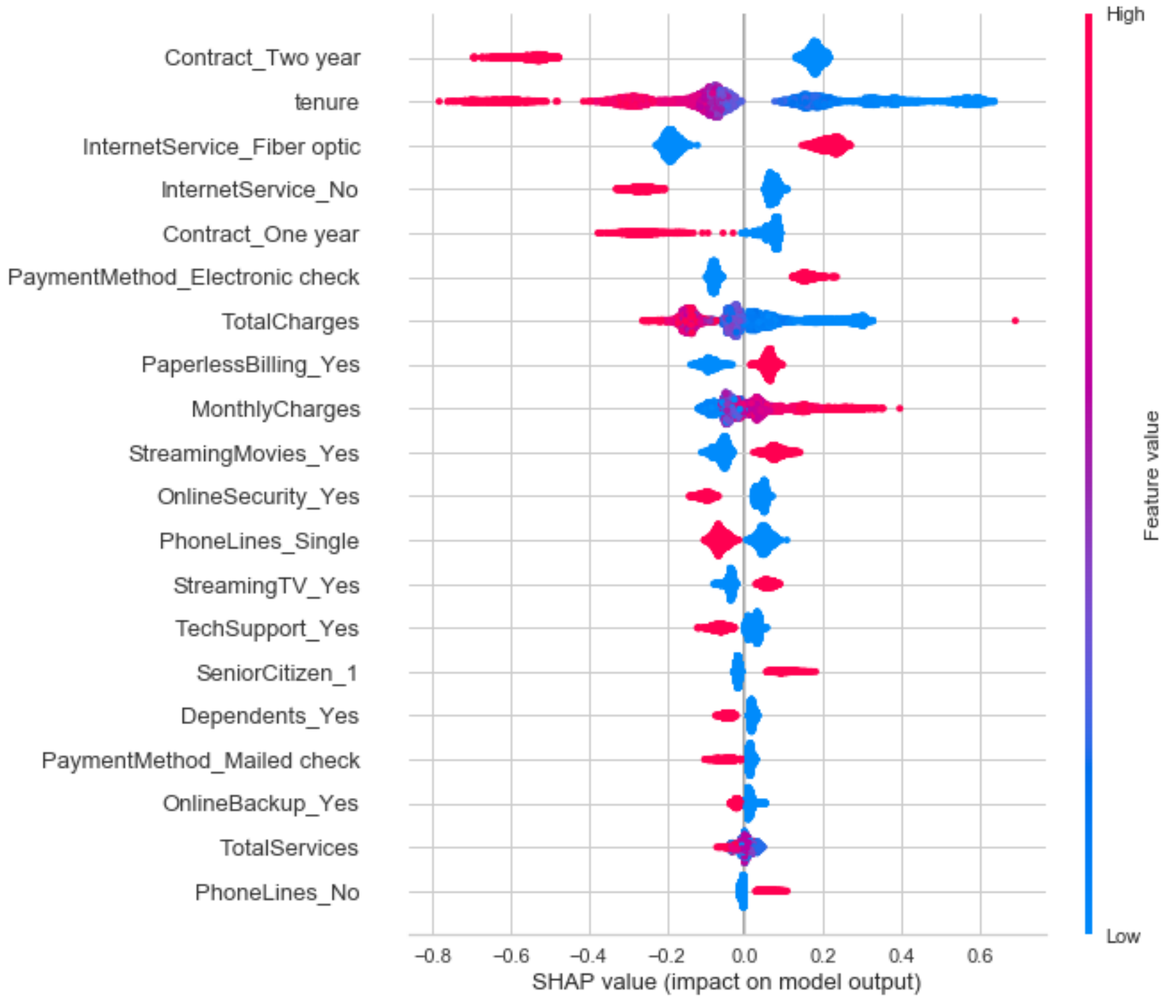
The output of LIME is a list of features, reflecting the contribution of each feature to the prediction of a data sample. This provides local interpretability, and it also allows to determine which feature changes will have most impact on the prediction.
The explainer so considers one observation at one time and shows not only the output of the classification but also what contributed the most and how.



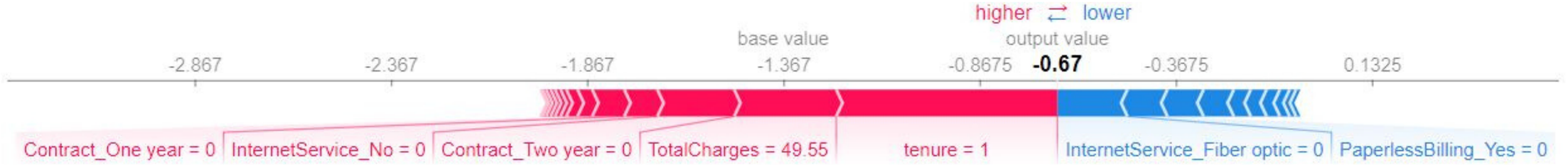
Shap Explainer

SHAP is a method to explain individual predictions.
The idea is using game theory to interpret target model.

The plot sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output.

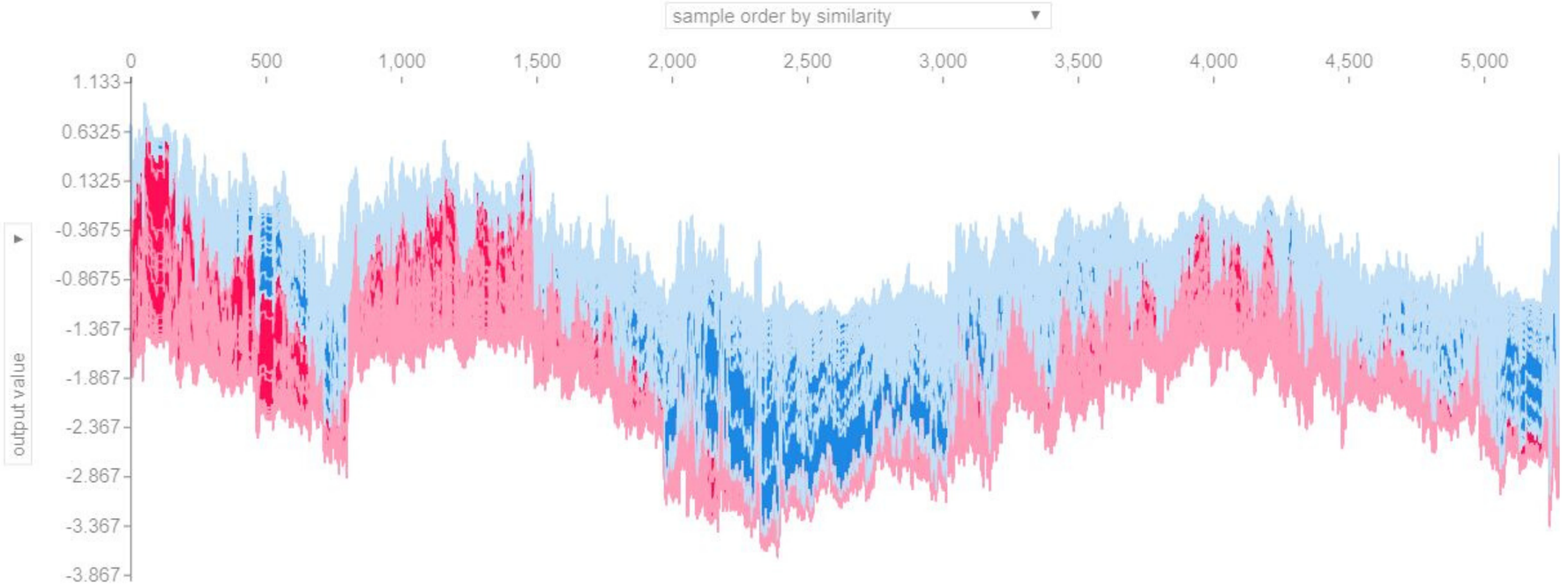


Shap Explainer



The above explanation shows how each feature contributes to the output value. Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue.

If we take many explanations such as the one shown above, rotate them 90 degrees, and then stack them horizontally, we can see explanations for an entire dataset



Remarks

From our model it is possible to obtain some insights:

- **Monthly costs:**
  The most important feature in the building of the GB model is "MonthlyCharges", followed by the highly correlated "TotalCharges". It means that cost per month is relevant in choosing a Telecommunication company. This have to be contestualized in today's market, there are different players and so price could be a discriminant since the service offered is usually replaceable.

- **Client history:**
  As shown also during the EDA, and confirmed by the and Shap plot, new clients are more likely to leave the company. In our model "Tenure" is a good predictor, what we get from this is that new clients need extra attention since they're more prone to change provider.

- **Type of contract:**
  Another good discriminant is the type of contract. Longer contract duration usually implies more probability to stay with the company. We suppose this could be because in longer contracts it's more likely to find contract termination clauses.

Conclusion & Possible Developments

Conclusion:
We suggest to use the model we built on new customers data in order to understand which clients are more likely to leave.
After that the optimal choice would be to focus at the value of those clients for the company (Customer lifetime value) and decide if the optimal choice would be to try to retain them.
In this case the company should leverage on the predictors that we have highlighted during the analysis and also shown through the explainers.

Further Developments:

- Augment the data with extra informations
- Extend the feature engineering process
- Explore possible clusters in the client base

THANK YOU FOR YOUR ATTENTION