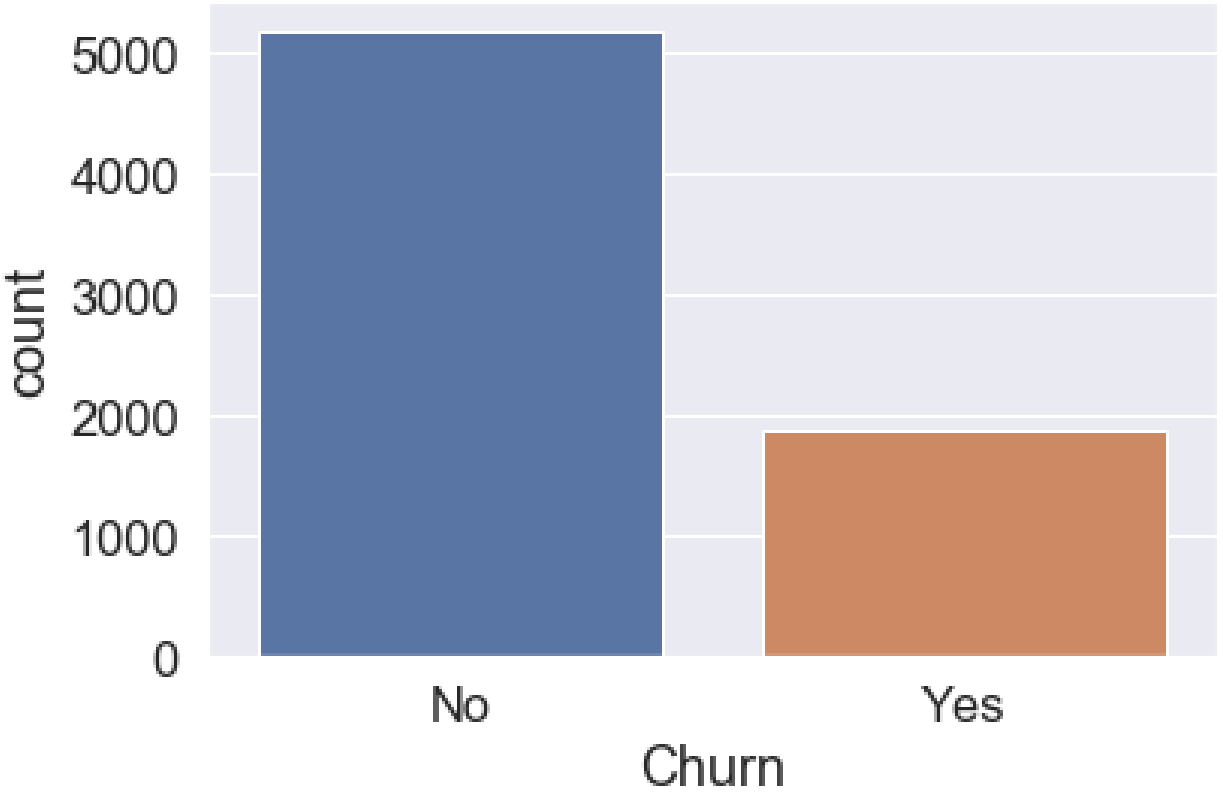
1. Data Overview

- In our data set there are 21 variable 20 of them are used as features for predicting customers churn (target variable)
- There are total 7043 observations
- The data was already preprocessed and contain no missing values
- As you can see from “Types of features” figure, TotalCharges is an “object” type, however it contains numerical values. That is why we convert it to float.

Types of features

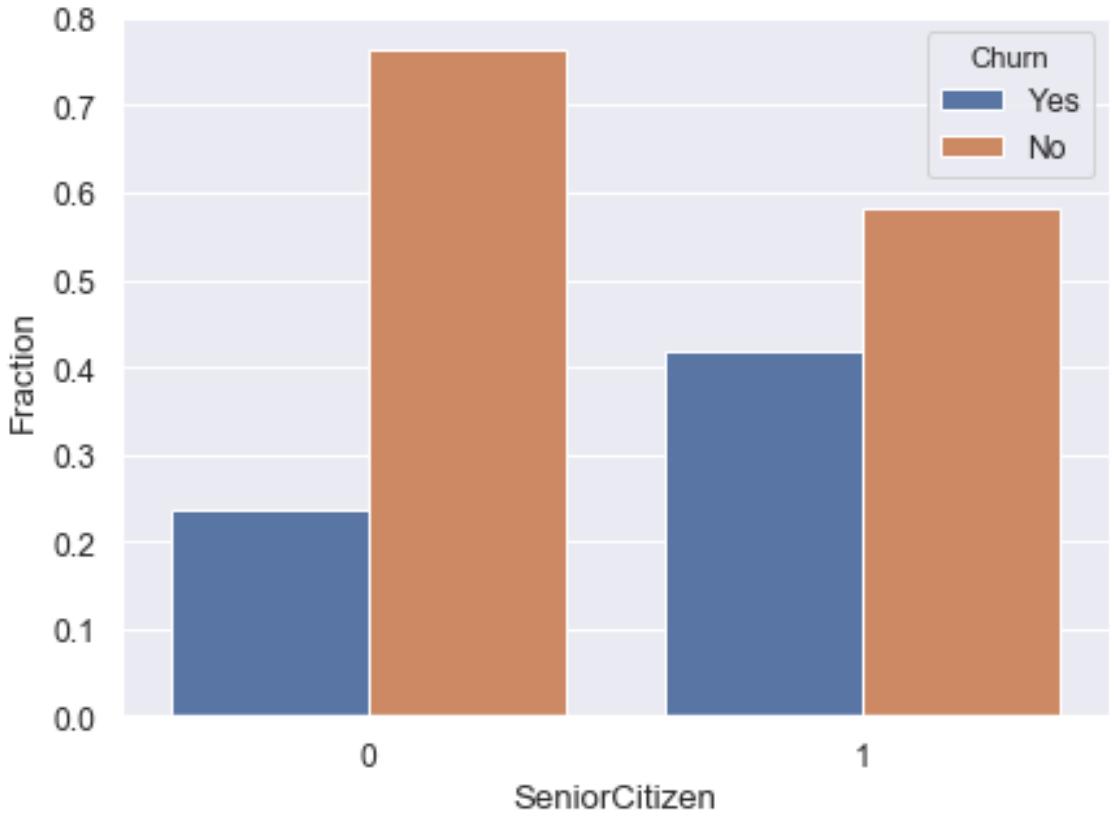
customerID	object
gender	object
SeniorCitizen	int64
Partner	object
Dependents	object
tenure	int64
PhoneService	object
MultipleLines	object
InternetService	object
OnlineSecurity	object
OnlineBackup	object
DeviceProtection	object
TechSupport	object
StreamingTV	object
StreamingMovies	object
Contract	object
PaperlessBilling	object
PaymentMethod	object
MonthlyCharges	float64
TotalCharges	object
Churn	object

2.1 Data Visualization



Our data contains more observations for customers that do not churn, thus the dataset is imbalanced and we have to balance it in order to create a better predictive model.

2.2 Data Visualization



This graph indicates that fraction of senior citizens that did churn is much higher that for those who were not senior citizens, which give us an idea about introducing better offers or discounts for senior citizens in order to retain them as customers.

3. Method Description

1) The categorical variables were converted to numerical via one hot encoding.

2) The imbalance of dataset was fixed through applying SMOTE method (which is a state of the art minority over-sampling technique).

3) The data was split into two groups: 70% of the data were in the training set and other 30% were in testing set.

4) For every model 5-fold stratified cross validation was applied in order to select the best model that would not overfit the data.

5) I used F1-score metric to compare different models as it represents a blended precision and recall metrics that are important when we are dealing with imbalanced datasets. In order to perform error analysis I trained model on different fractions of training data and plotted the results for F1-score vs fraction of available training data.

6) Also I removed highly correlated variables from the dataset and standardize the data.

7) In our data there we 1869 instances of customers who churned and 5174 instances of customers who did not churn. Thus, if we predict that everybody did not churn we would achieve about 73% accuracy, but zero F1-score.

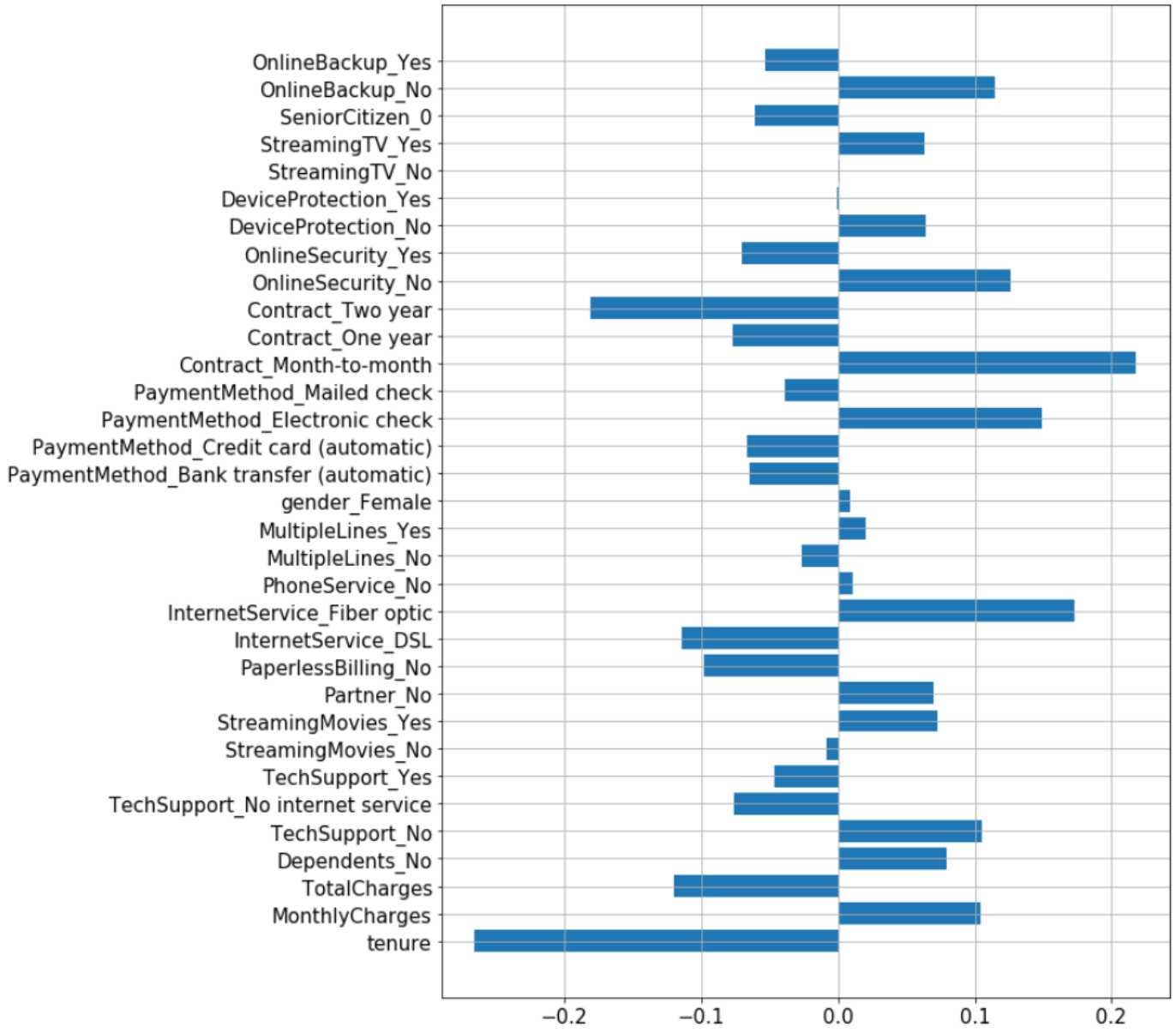
8) Logistic Regression and XGBoost algorithms performed almost the same giving about 0.62 value for F1-score and for our testing set. Accuracy score for XGBoost was 0.78 whereas for Logistic Regression it was 0.75. I also tried SVM with both linear and Gaussian kernels, but it performed poorer than other models in terms of F1-score (even though the accuracy score was slightly better than for Logistic Regression model.

4. Feature Importance

From the information contained in the chart we can establish the following relationship between predictors and target variable:

- The longer the tenure period, the less likely the customer will churn (negative coefficients for tenure and Contract_Two year features and positive for Contract_Month-to-month)
- The more price our customers pay per month, the more is the probability that they will churn (positive MonthlyCharges coefficient).

We will use this connections when we suggest the recommendations on how to reduce the customer churn.



5. Recommendations for Reducing Churn

Based on the barplots for relative counts of customers across different groups I would suggest the following recommendations:

- In both Logistic Regression and XGBoost models features Tenure and MonthlyCharges were important, so we should try to affect them. From the violin plots for tenure and monthly charges we can see, than on average people with smaller tenure period and higher monthly charges tend to churn more. Thus, we can introduce an decremental payment plan where the customer's payment would decrease with the time. In this case our customers will have a good incentive to stay longer with our company and pay less. This change should reduce the churn according to our analysis,
- Improve technical support. It can be reached for example through hiring more people, making it easier for people to have technical support – introduce new options such as online chat, phone call, video call, in-person help.
- Better explain the importance of device protection and online security as people who neglected security issues tended to churn more.