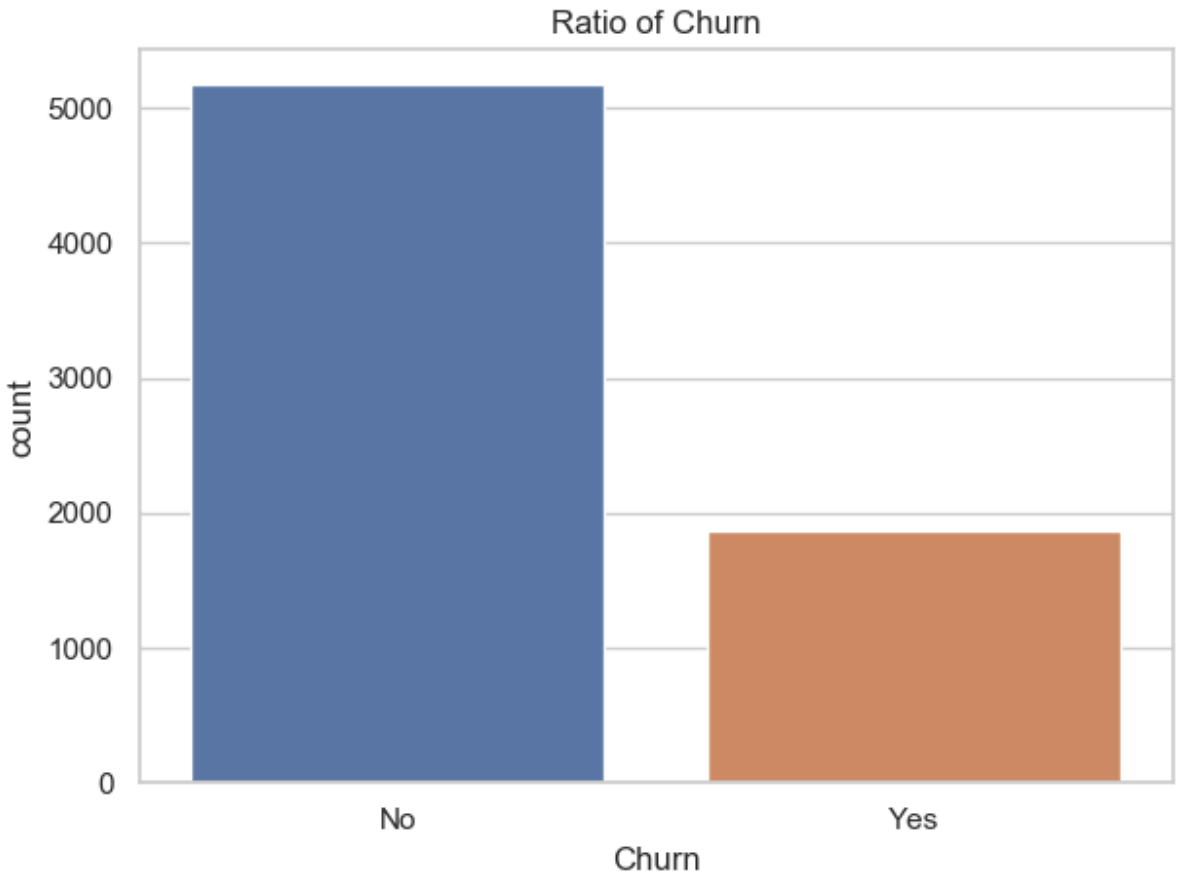
Prediction of Churn

By Aleksandra Pereverzeva

Churn Ratio



1869 (26.53%) of users churned, while 5174 (73.46%) of users did not churn.

Distribution of Tenure

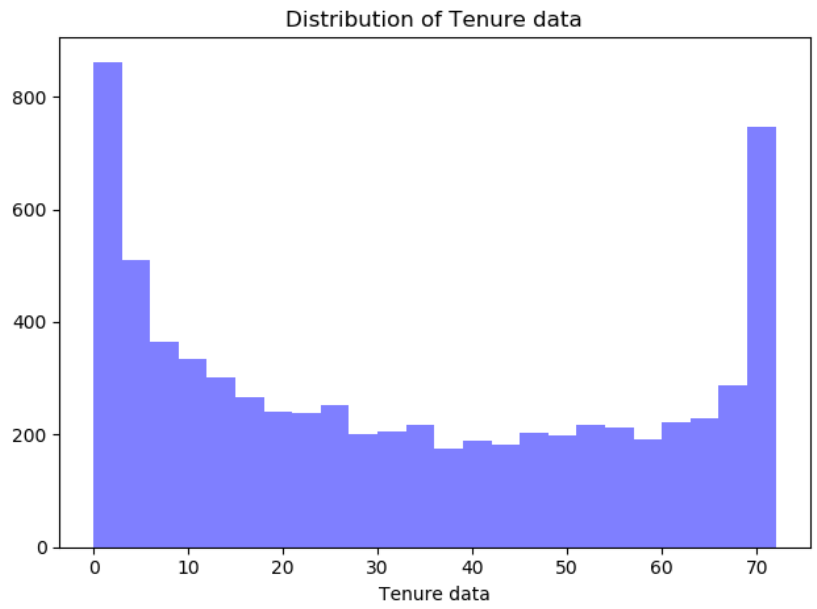


Figure 1 Distribution of Tenure data

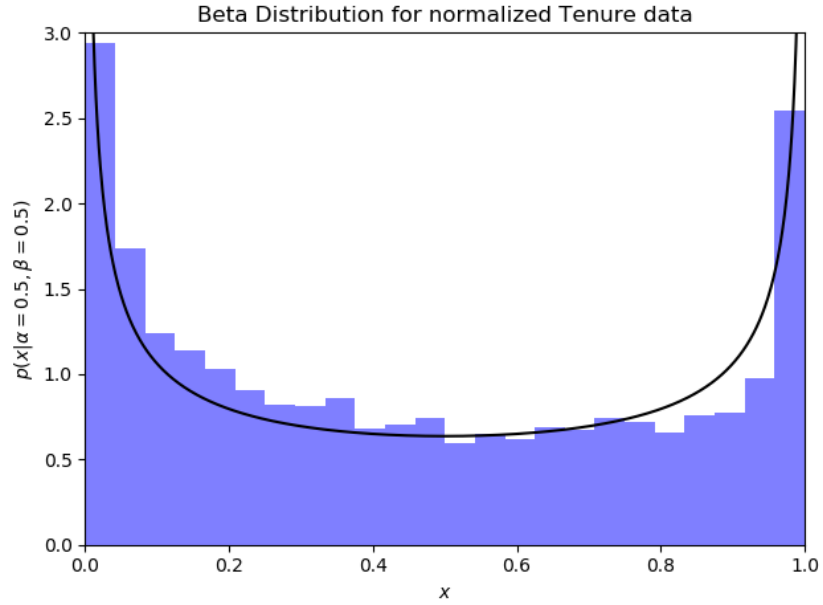
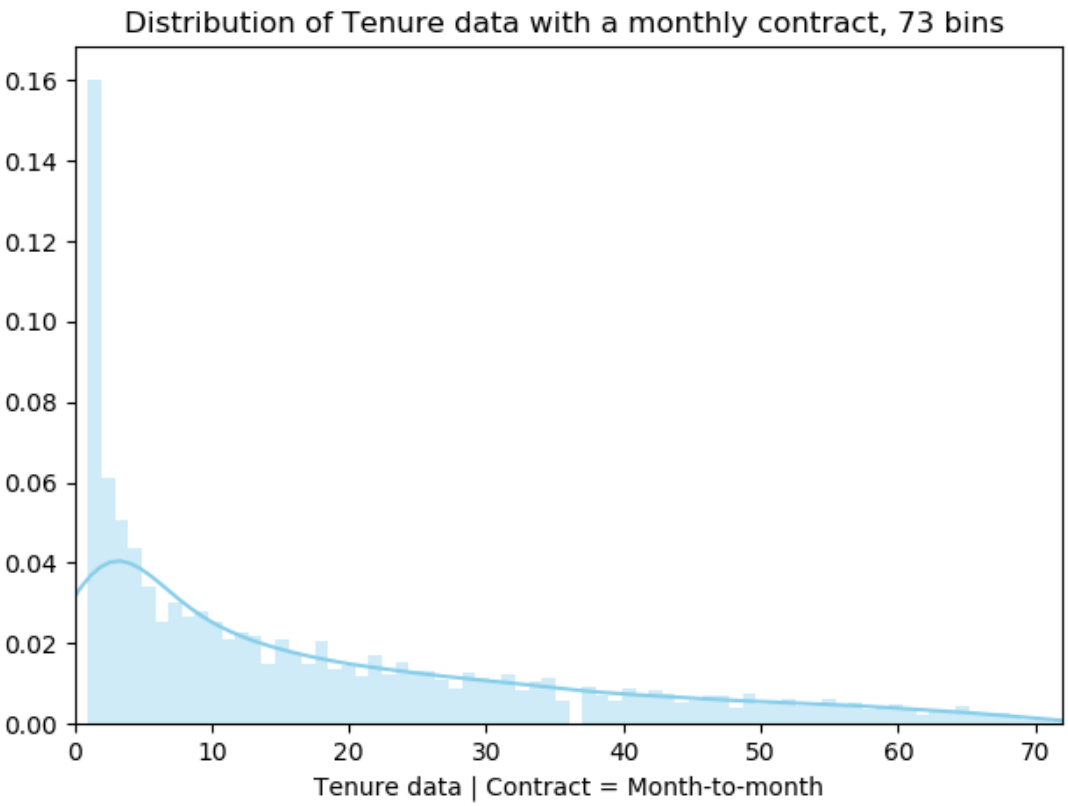


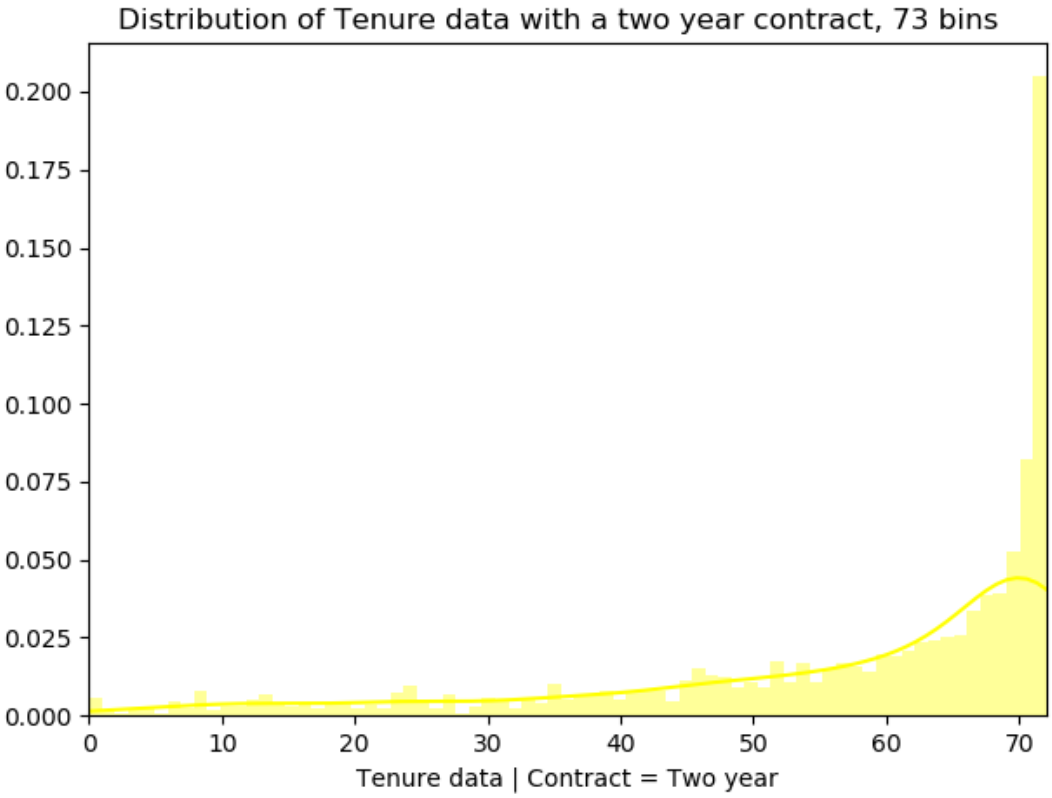
Figure 2 Tenure data with fitted Beta distribution

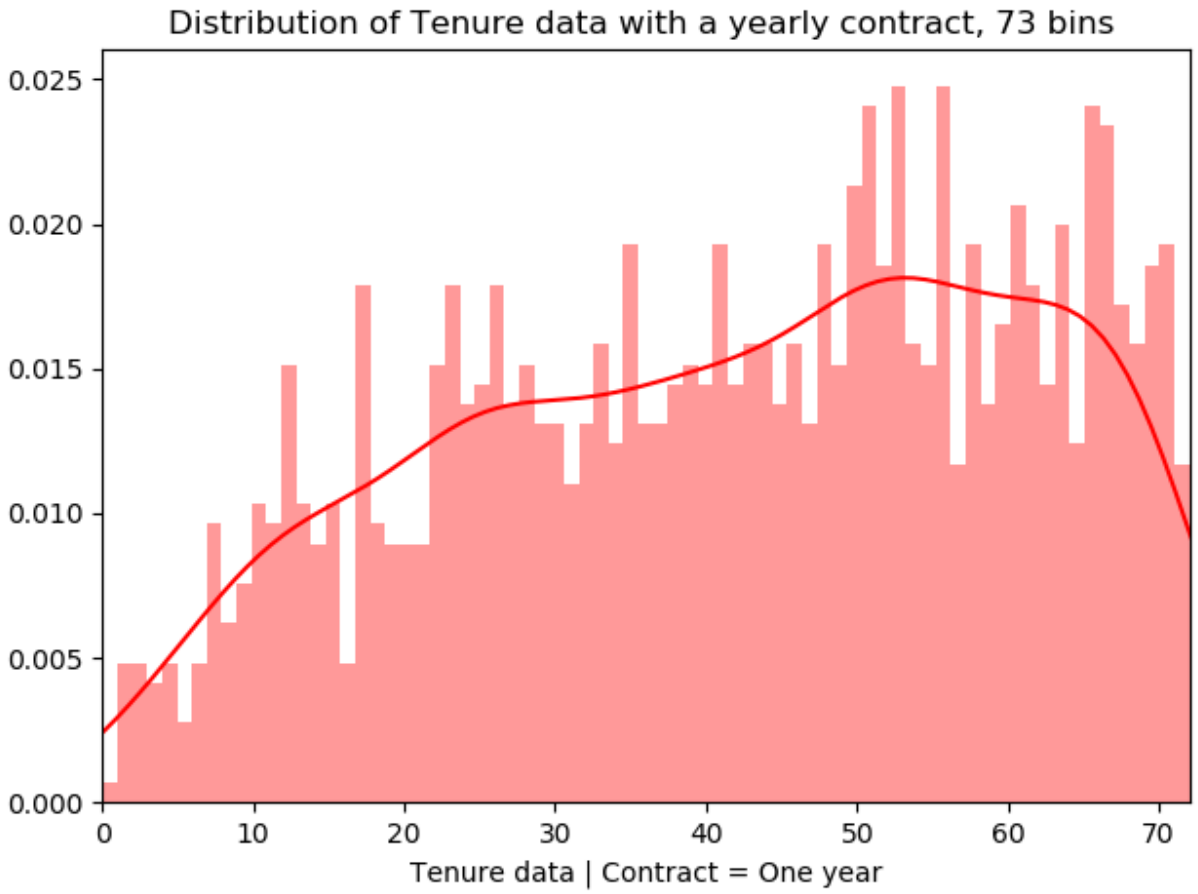
The beta distribution does not fit perfectly but close enough. I believe this is the true distribution of tenure data. This tells us, that most of the customers tend to have either a very long or a very short history of cooperation.

ARPU

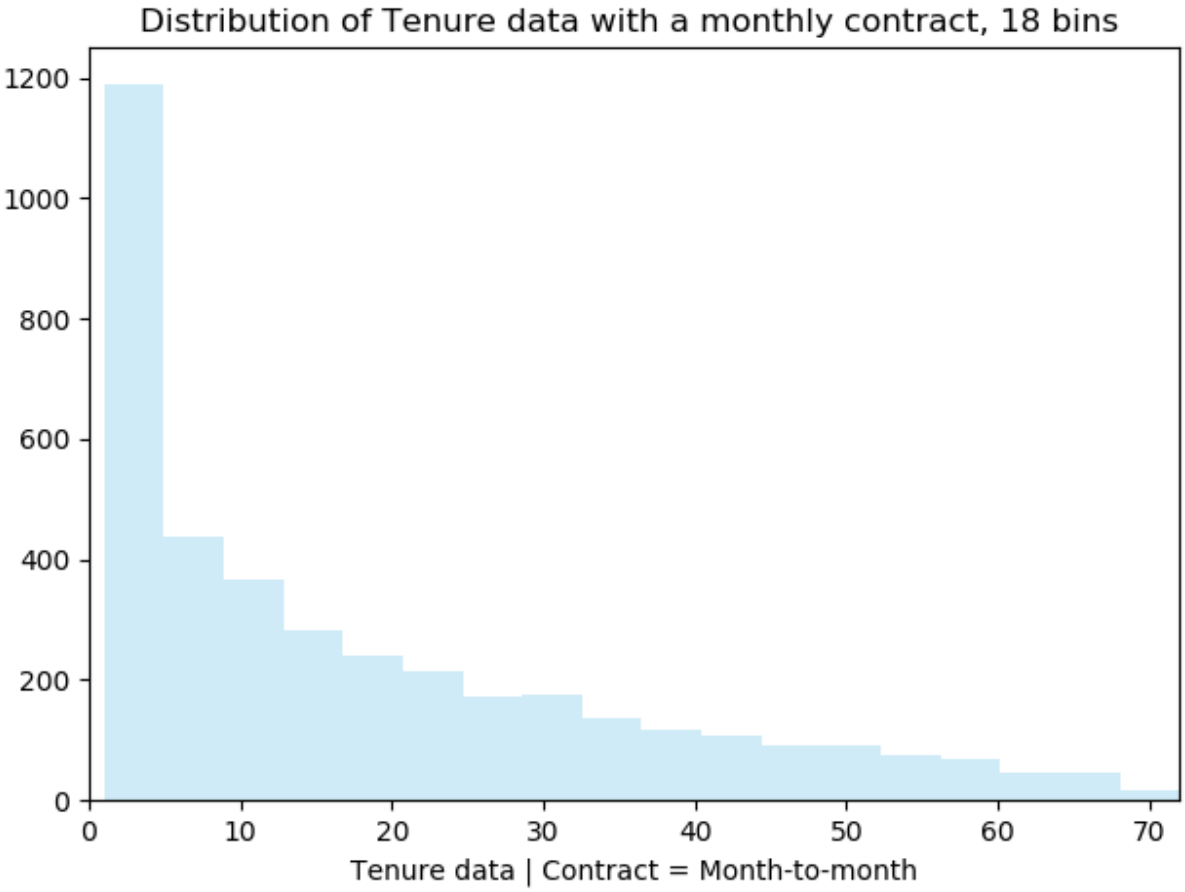
There are three types of the contract: month-to-month, one year and two year. First it might be interesting to have a look on the data not of the tenure divided by the contract type:

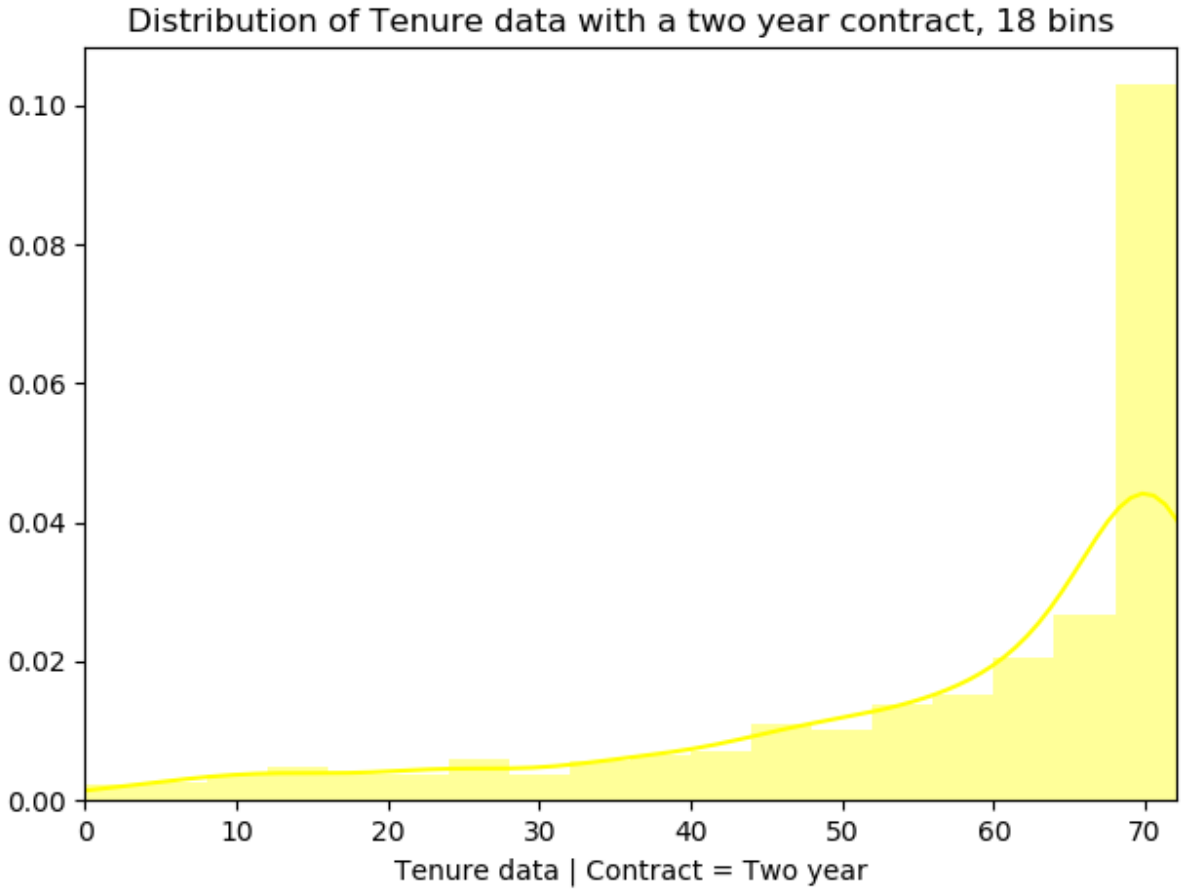

Distribution of Tenure data with a monthly contract, 73 bins
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
70
Tenure data | Contract = Month-to-month

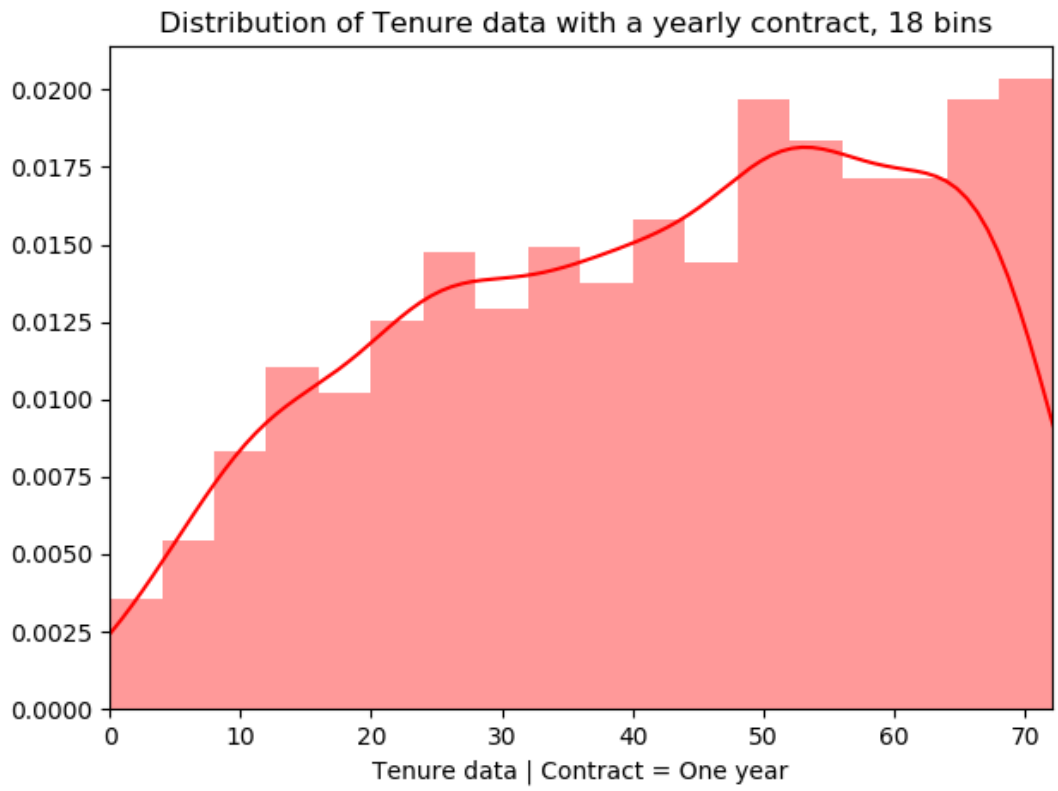

Distribution of Tenure data with a two year contract, 73 bins
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
50
60
70
Tenure data | Contract = Two year



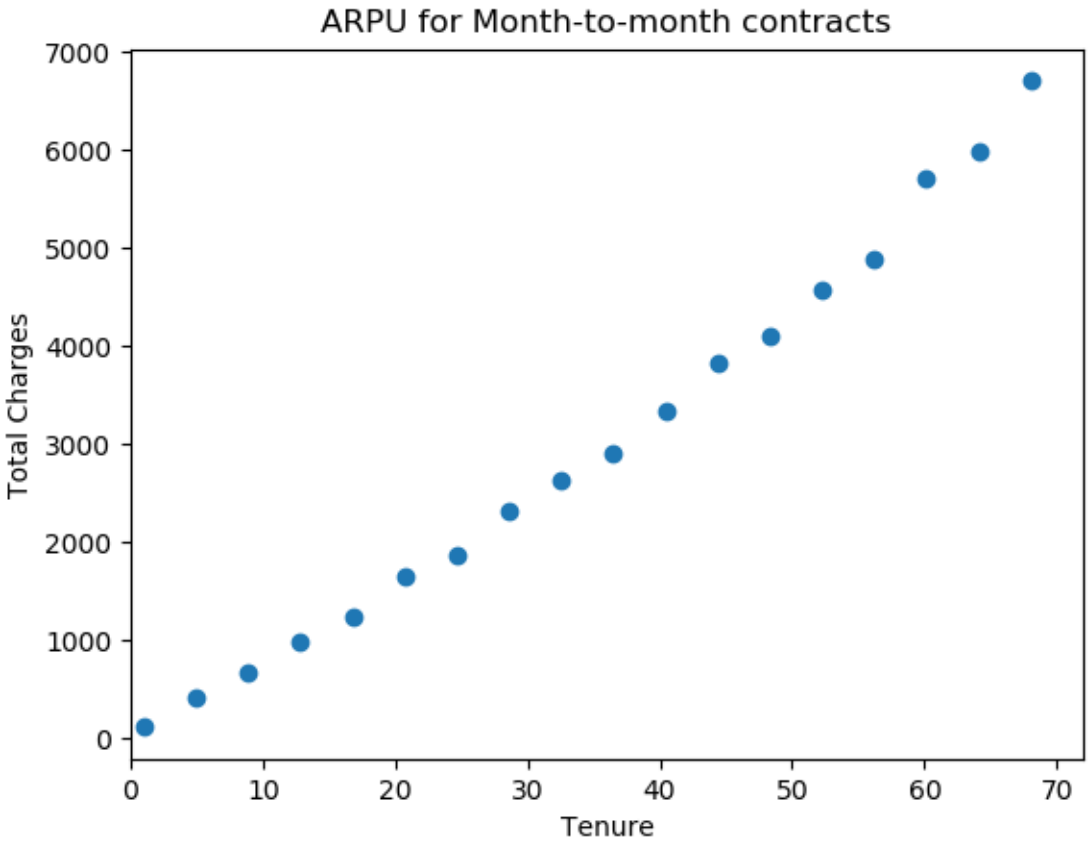
In this case all the individual values of tenure received the bin of their own, we can smoothen the graphs by assigning ~4 values to one bin:

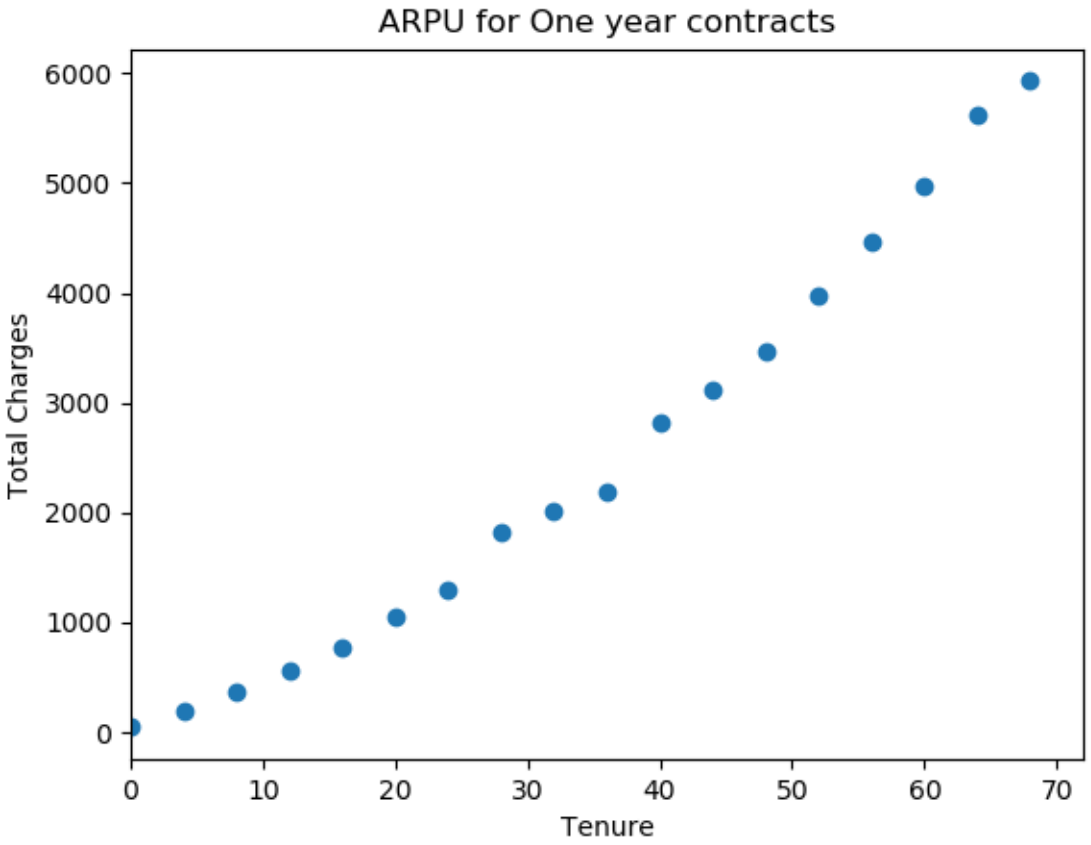

Distribution of Tenure data with a monthly contract, 18 bins
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
Tenure data | Contract = Month-to-month

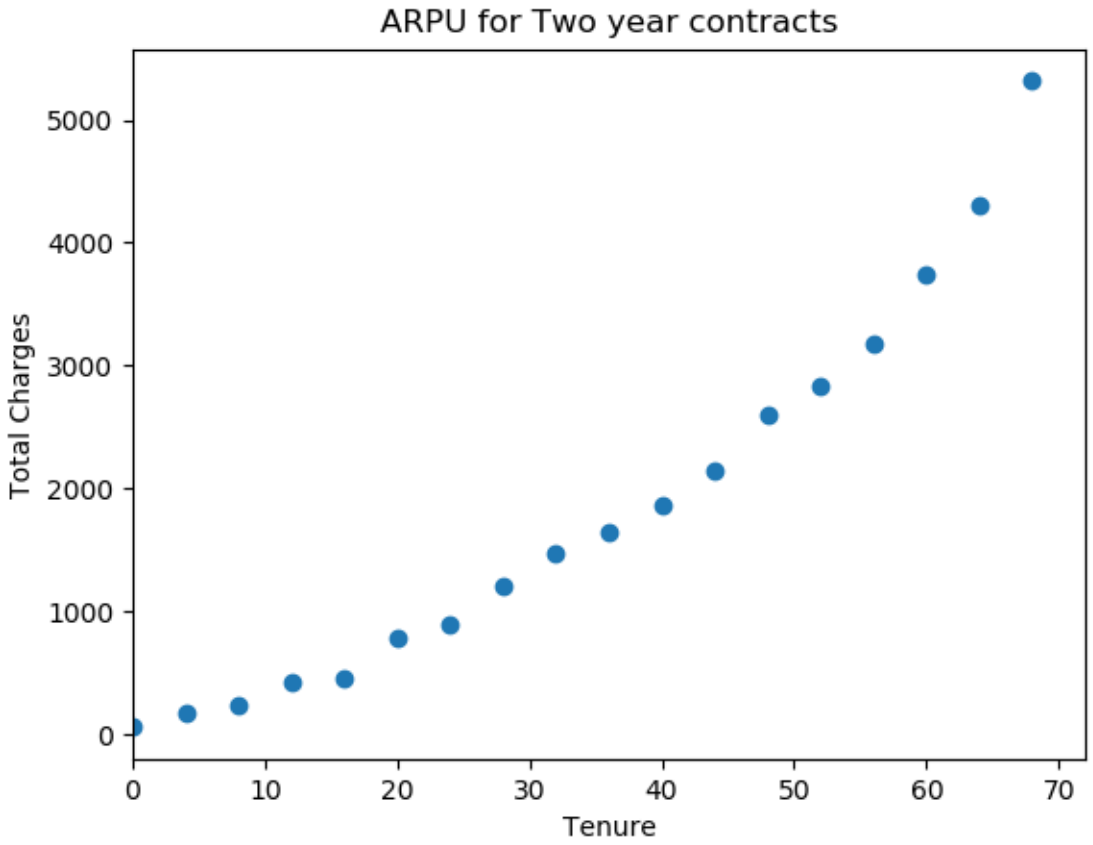

Distribution of Tenure data with a two year contract, 18 bins
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
70
Tenure data | Contract = Two year



As we can see it is a very favorable bin width: the graphs look smoother, but the appropriate tendencies are still visible. Now the corresponding ARPU graphs can be shown:


ARPU for Month-to-month contracts
Total Charges
Tenure
0
1000
2000
3000
4000
5000
6000
7000
0
10
20
30
40
50
60
70


ARPU for One year contracts
Total Charges
Tenure
0
1000
2000
3000
4000
5000
6000
0
10
20
30
40
50
60
70



The tendencies on the graphs show us that the average revenue per customer is increasing directly proportionally with the value of tenure. This means that the customers with the longer history are more valuable for the company despite the contract types.

The most profitable churned customers

It's hard to draw the line between the most profitable and the less profitable customers. For a sake of the experiment let us consider that the most profitable customers make 10% of all the customers that churned.

That can give us an approximate portrait of the most profitable customer that churned:

Gender	Male
SeniorCitizen	0
Partner	Yes
Dependents	No
Tenure	66
PhoneService	Yes
MultipleLines	Yes
InternetService	Fiber optic
OnlineSecurity	No
OnlineBackup	Yes
DeviceProtection	Yes
TechSupport	No
StreamingTV	Yes
StreamingMovies	Yes
Contract	One year
PaperlessBilling	Yes
PaymentMethod	Electronic check

Monthly Charges mean: 101.85000000000001.

Total Charges mean: 6029.261827956989.

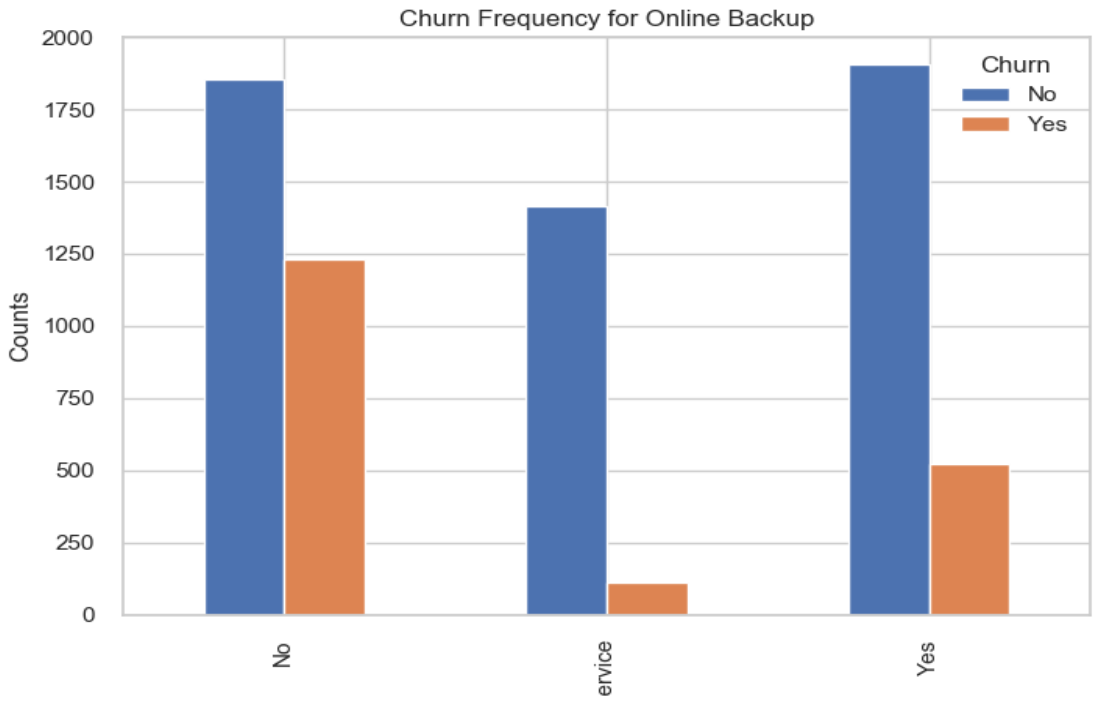
Probably the most important thing that is visible that such customers use Fiber optic for the Internet access and Electronic checks as a payment method. This might be useful for the creating of the model.

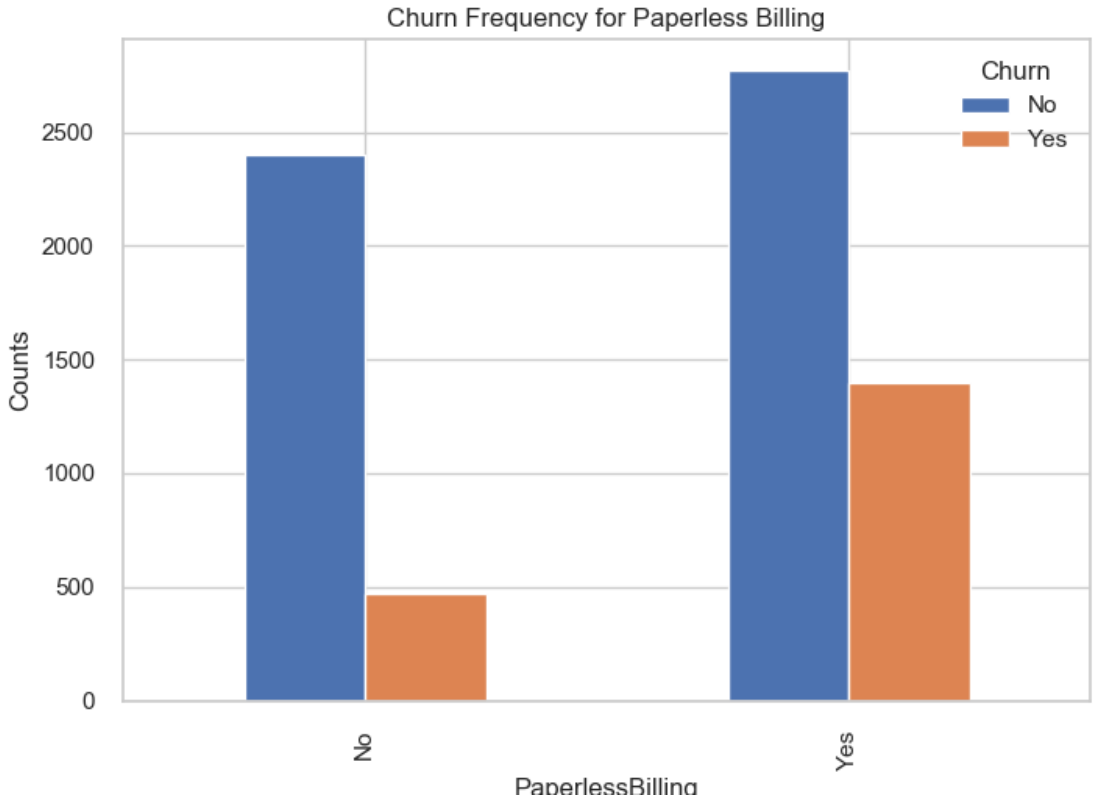
Churn main drivers

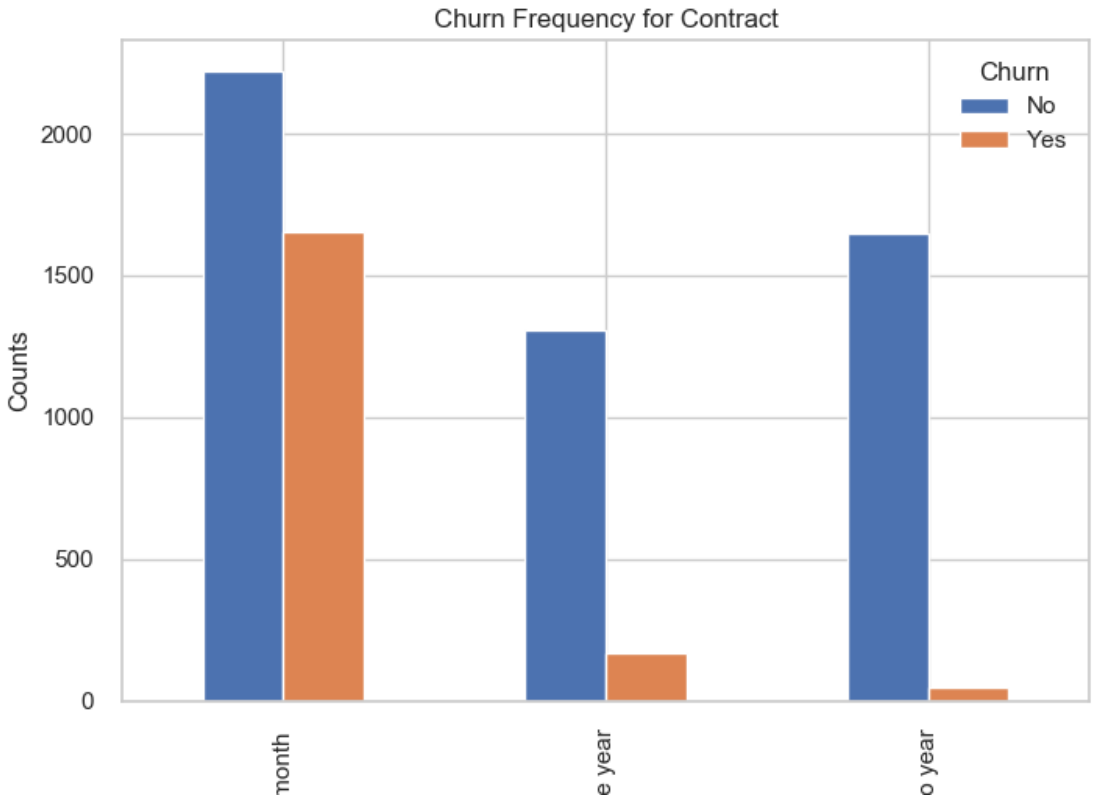
There are different methods for identifying the drivers of one category.

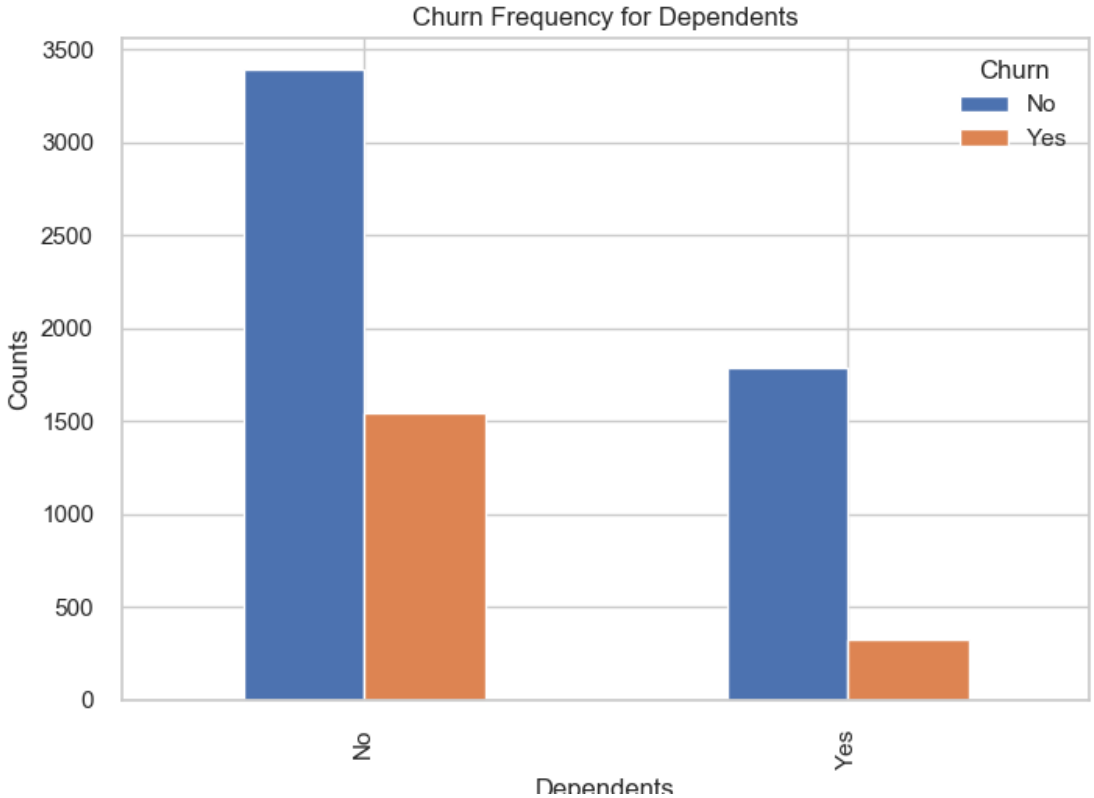
The first one is the most visually pleasing – graphs.

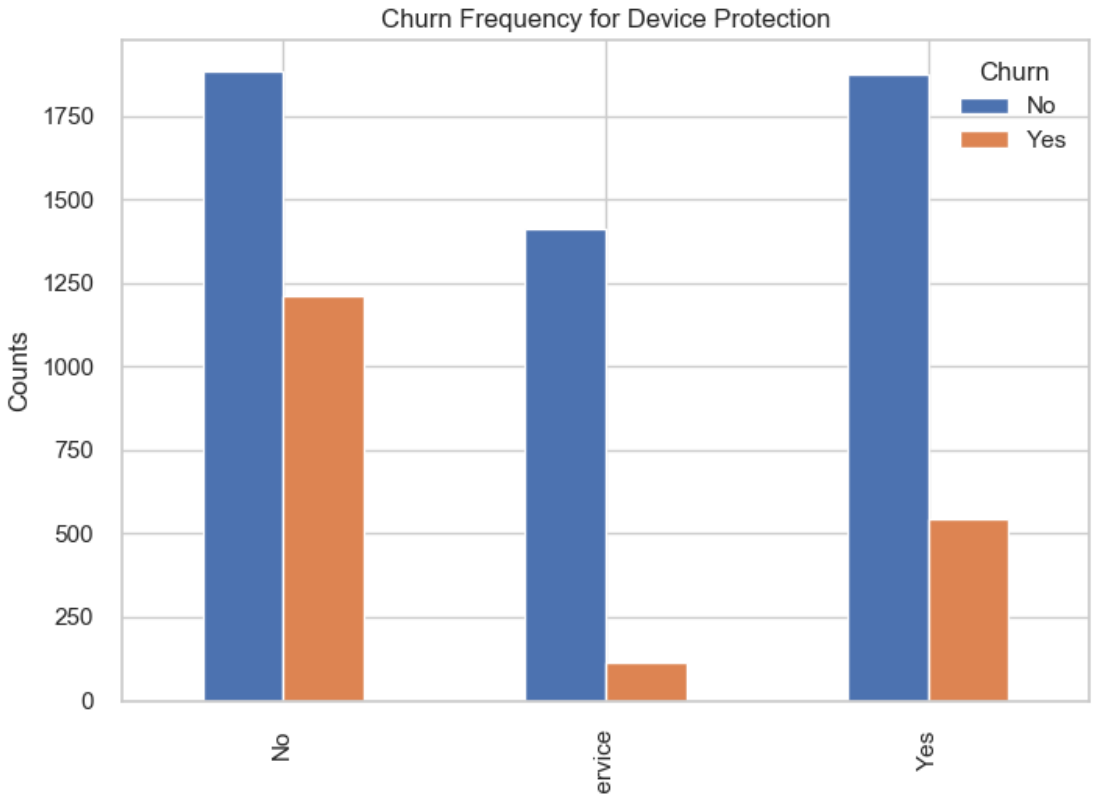
First the variables where the difference churned and not churned is drastic:

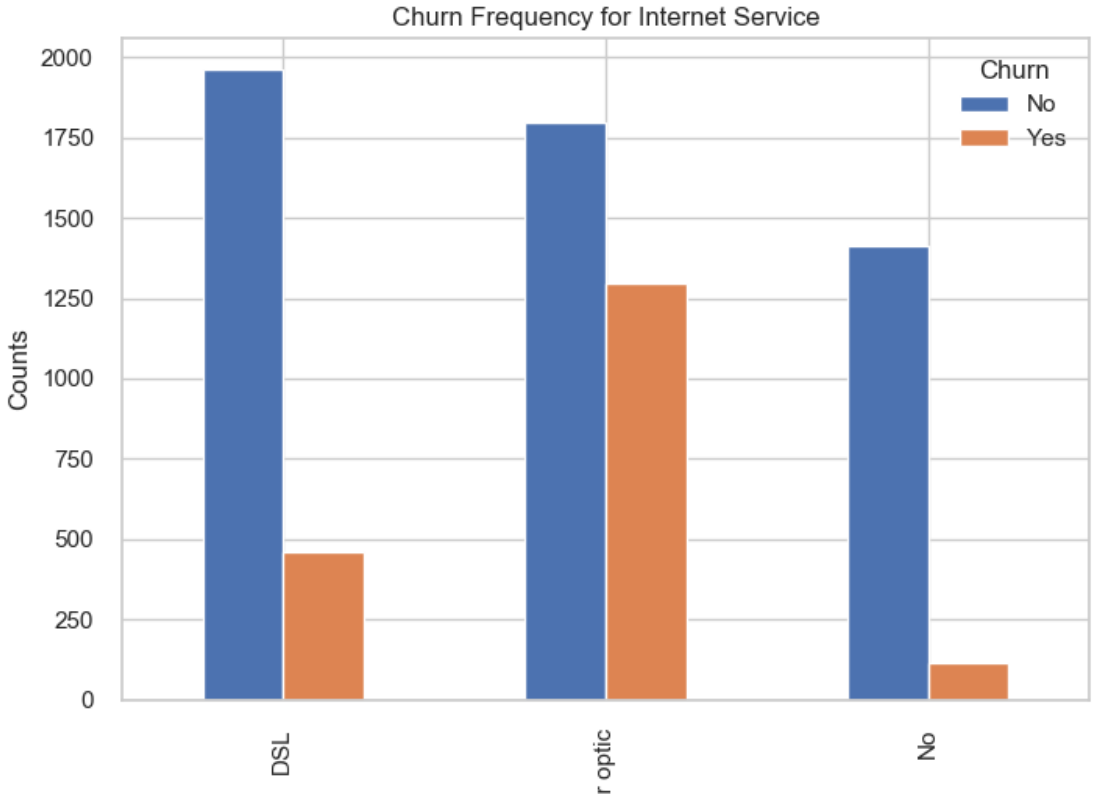

Churn Frequency for Online Backup
Churn
No
Yes
Counts
2000
1750
1500
1250
1000
750
500
250
0
No
ervice
Yes

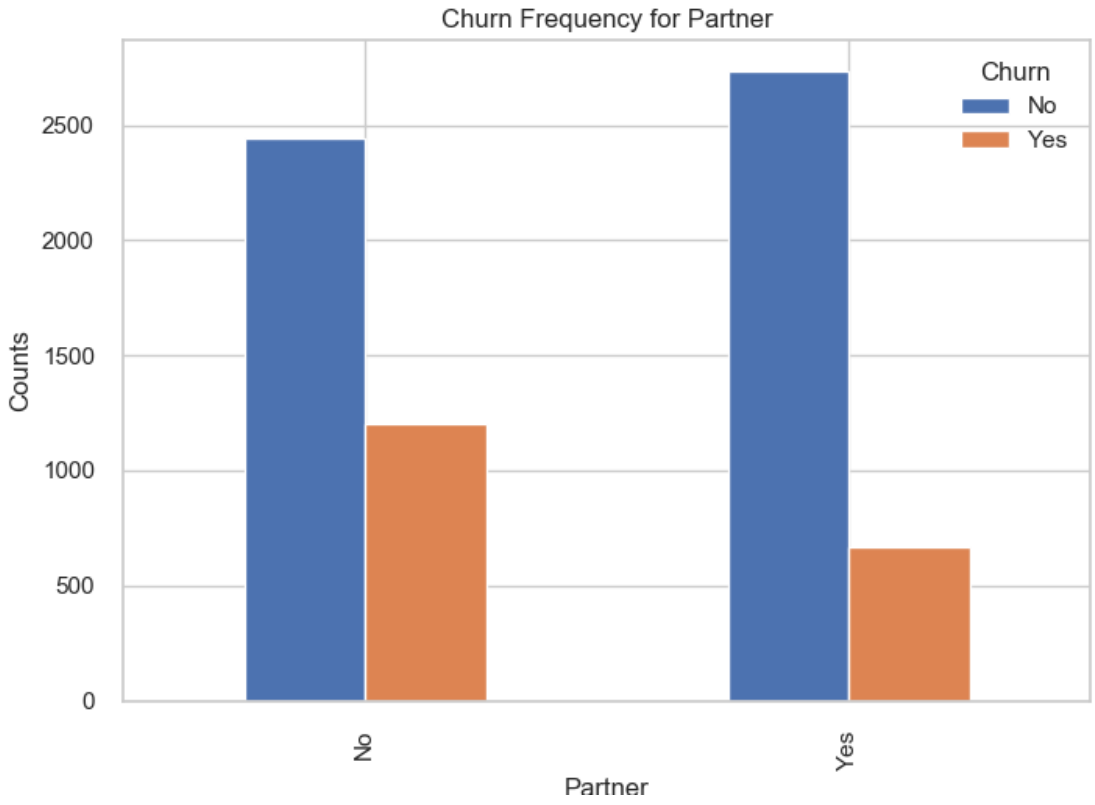

Churn Frequency for Paperless Billing
Churn
No
Yes
Counts
2500
2000
1500
1000
500
0
No
Yes
PaperlessBilling

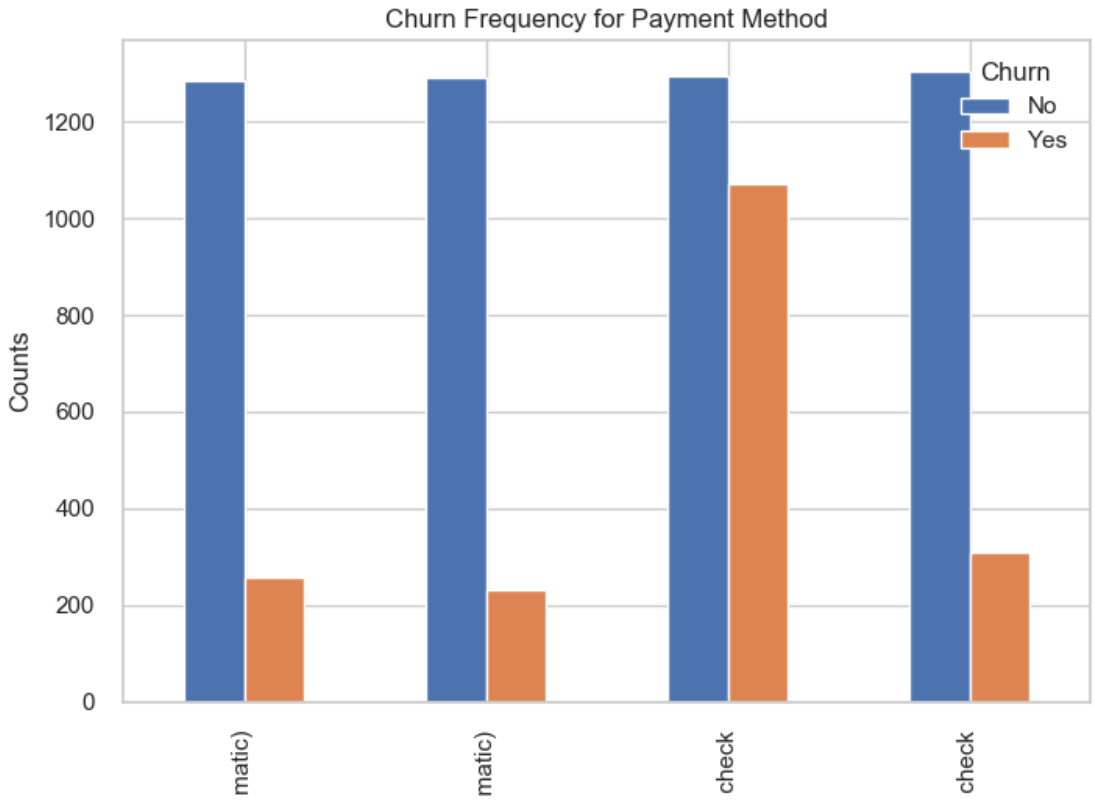

Churn Frequency for Contract
Churn
No
Yes
Counts
2000
1500
1000
500
0
month
e year
o year

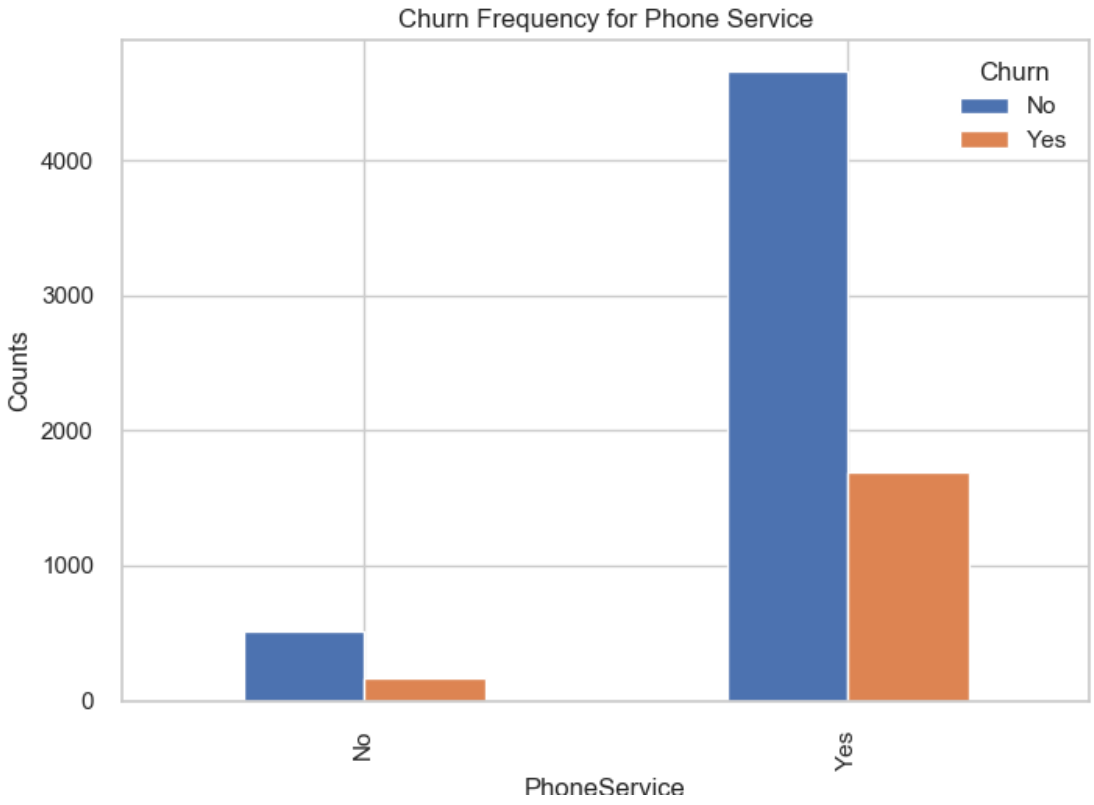

Churn Frequency for Dependents
Churn
No
Yes
Counts
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Dependents

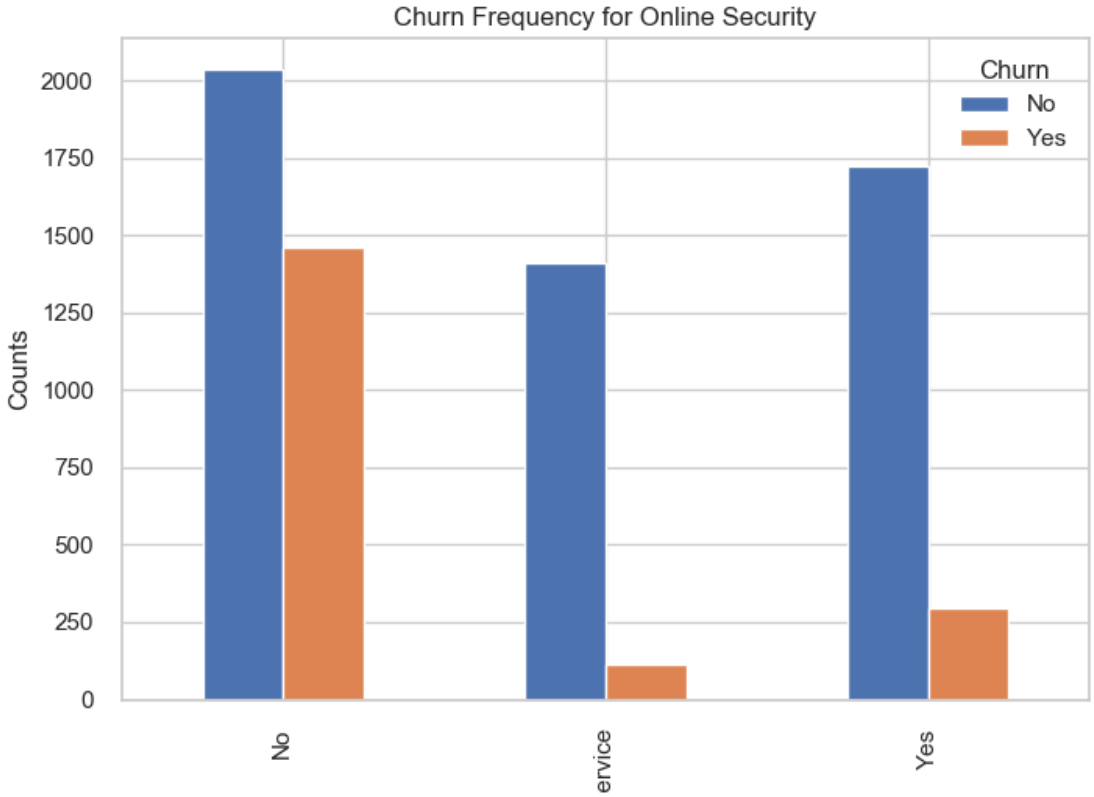

Churn Frequency for Device Protection
Churn
No
Yes
Counts
0
250
500
750
1000
1250
1500
1750
No
ervice
Yes

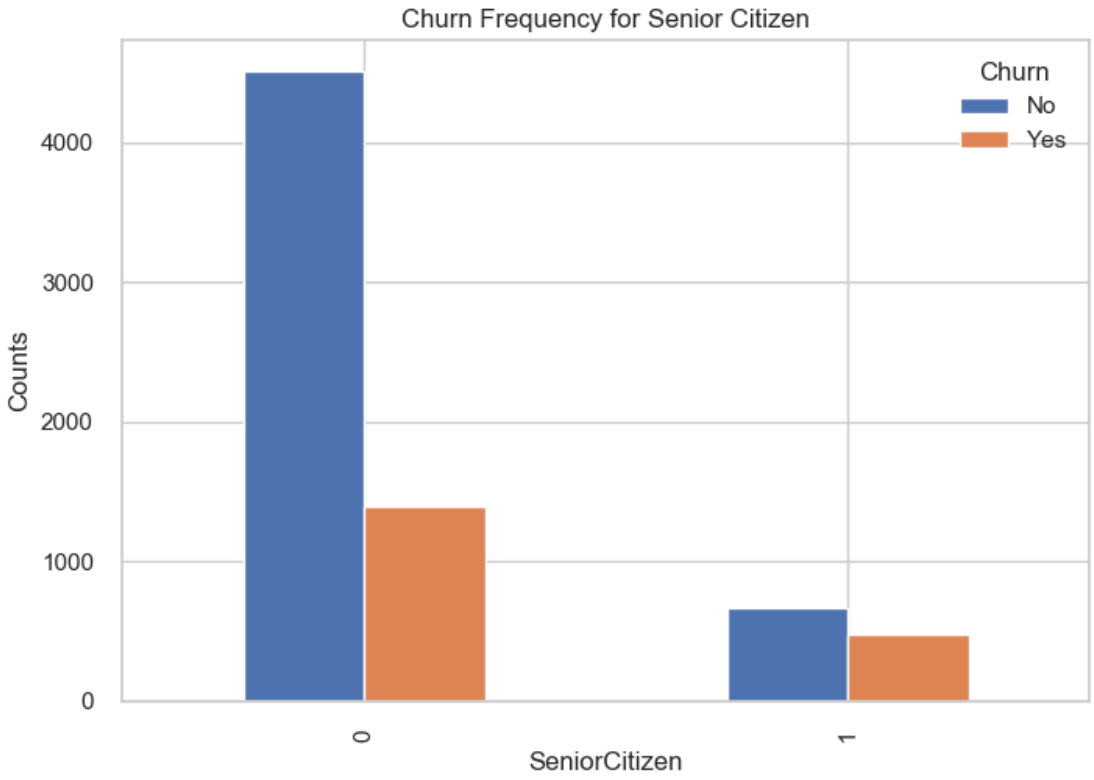

Churn Frequency for Internet Service
Churn
No
Yes
Counts
0
250
500
750
1000
1250
1500
1750
2000
DSL
r optic
No

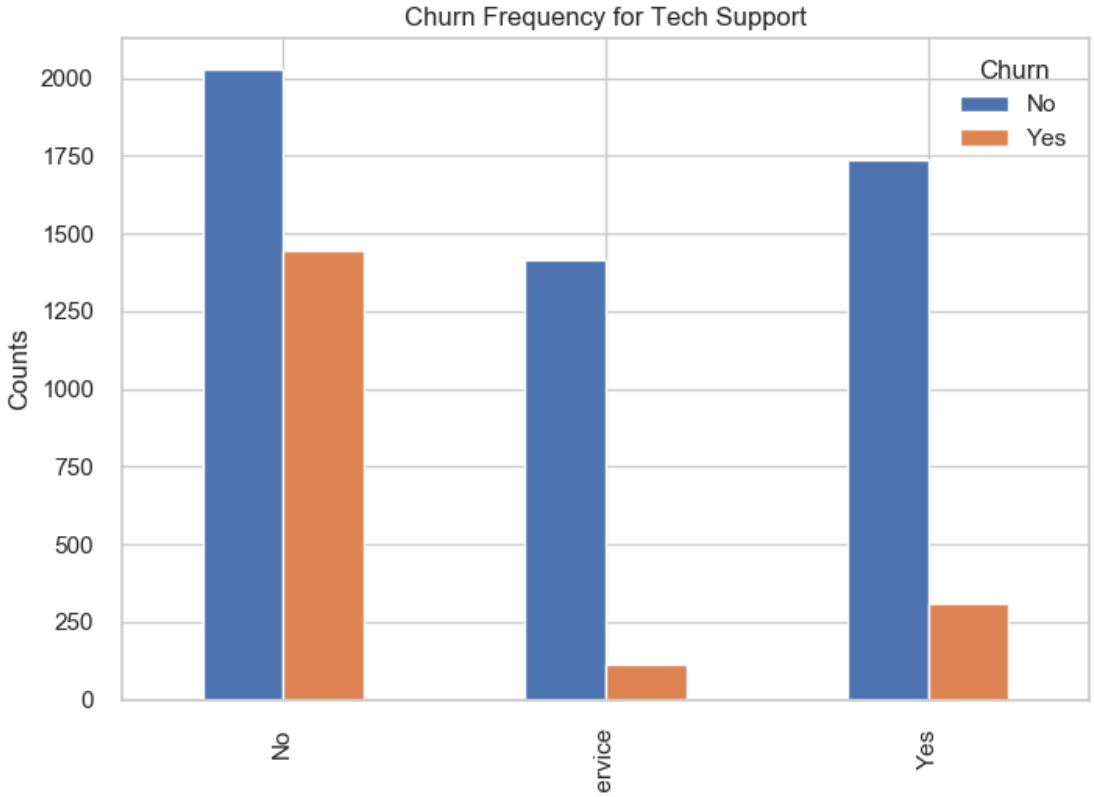

Churn Frequency for Partner
Churn
No
Yes
Counts
2500
2000
1500
1000
500
0
No
Yes
Partner

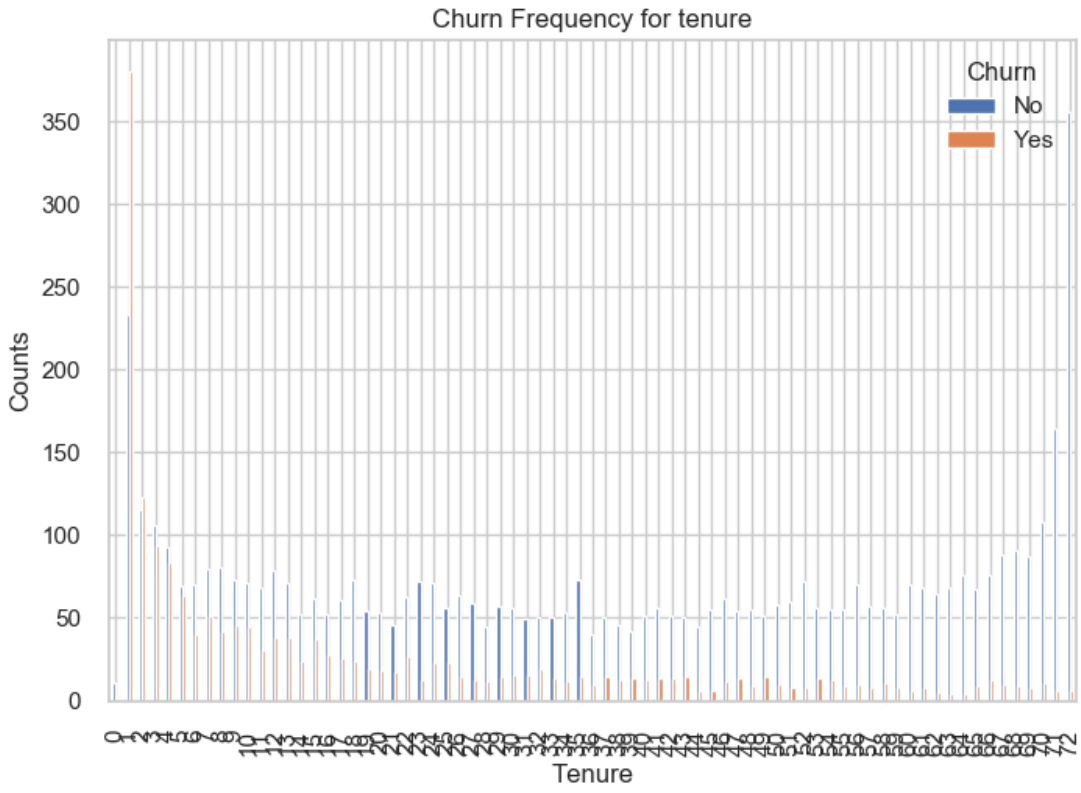

Churn Frequency for Payment Method
Churn
No
Yes
Counts
1200
1000
800
600
400
200
0
matic)
matic)
check
check

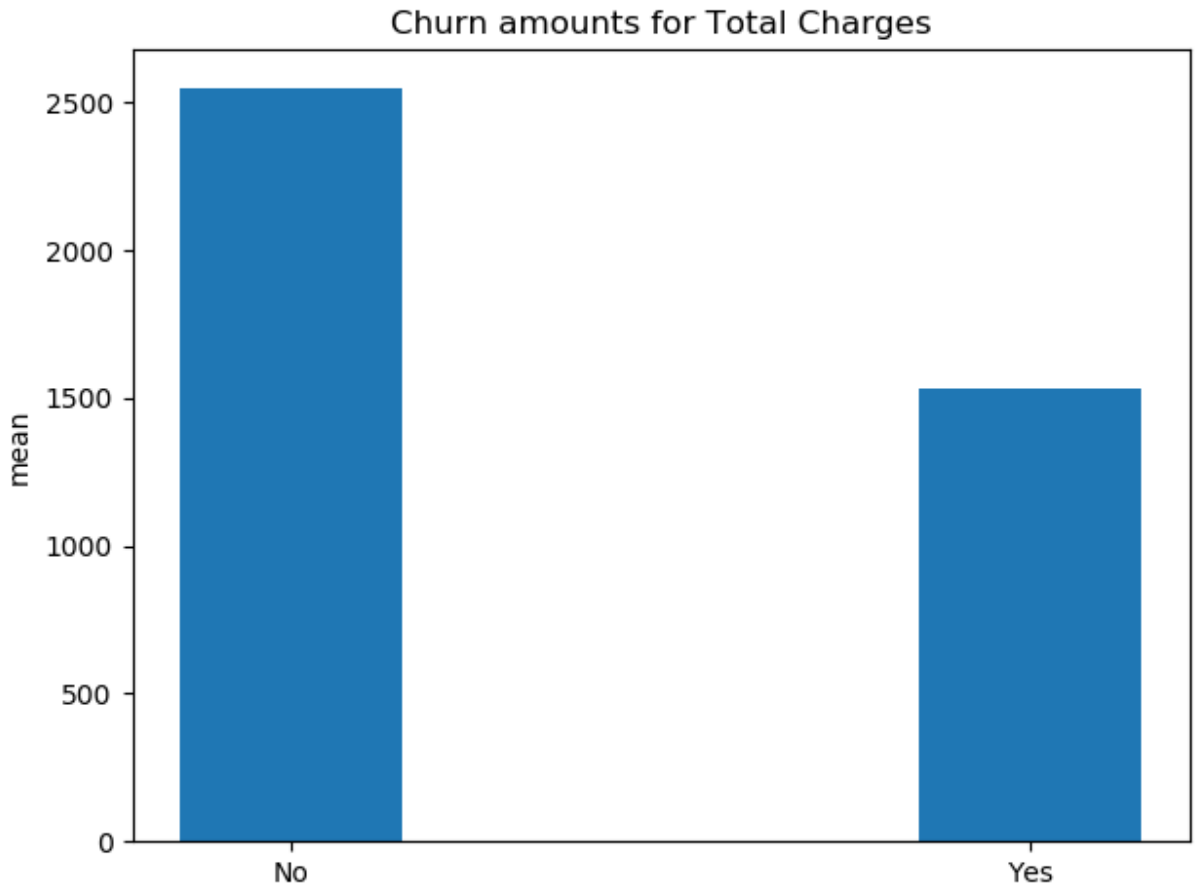

Churn Frequency for Phone Service
Churn
No
Yes
Counts
4000
3000
2000
1000
0
No
Yes
PhoneService


Churn Frequency for Online Security
Churn
No
Yes
Counts
2000
1750
1500
1250
1000
750
500
250
0
No
ervice
Yes


Churn Frequency for Senior Citizen
Churn
No
Yes
Counts
4000
3000
2000
1000
0
0
1
SeniorCitizen


Churn Frequency for Tech Support
Churn
No
Yes
Counts
2000
1750
1500
1250
1000
750
500
250
0
No
ervice
Yes


Churn Frequency for tenure
Churn
No
Yes
Counts
0
50
100
150
200
250
300
350
Tenure



On this graphs we can see that younger customers that do not history with the company, without Online Backup, dependents, device protection, partners, tech support or online security, but with phone service, that use paperless billing and electronic checks for payment and fiber optic for the internet with higher total charges are more likely to churn. All the features presented above can be drivers of churn.

According to the **Recursive Feature Elimination**, the drivers of churn are the following:

'SeniorCitizen_0', 'Partner_1', 'PhoneService_1', 'OnlineBackup_No',

'OnlineBackup_No internet service', 'InternetService_DSL',

'InternetService_Fiber optic', 'InternetService_No',

'OnlineSecurity_No', 'OnlineSecurity_No internet service',

'DeviceProtection_No internet service', 'TechSupport_No',

'TechSupport_Yes', 'StreamingTV_No', 'StreamingTV_No internet service',

'Contract_Month-to-month', 'Contract_Two year', 'PaperlessBilling_0',

'PaymentMethod_Bank transfer (automatic)',

'PaymentMethod_Credit card (automatic)'

(The 0 means, that the customer is not or does not have this feature.)

According to the logistic regression, the most valuable features and the corresponding weights are the following:

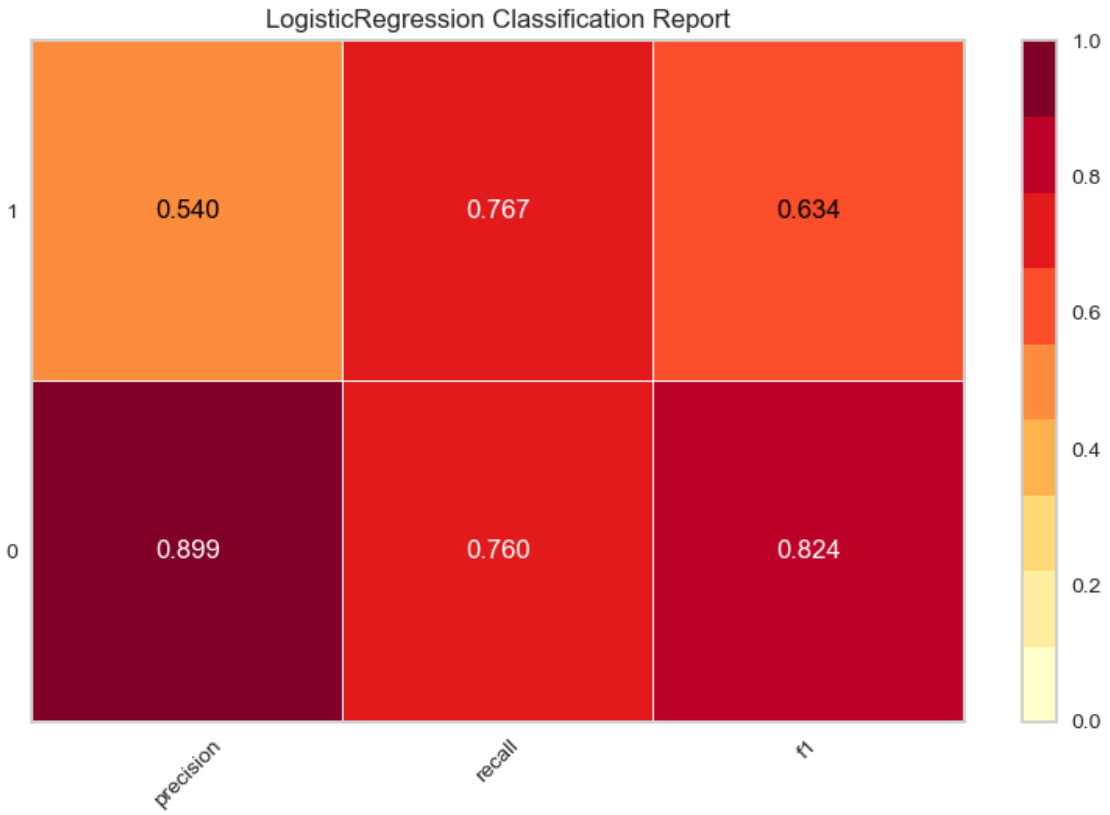
Contract_Month-to-month	0.525021
InternetService_Fiber optic	0.31268
PaymentMethod_Electronic check	0.257112
OnlineSecurity_No	0.220567
TechSupport_No	0.216618
PhoneService_0	0.132909
MultipleLines_No service	0.132909
PaperlessBilling_1	0.105837
OnlineBackup_No	0.090408
StreamingMovies_Yes	0.080112
StreamingTV_Yes	0.076018
SeniorCitizen_1	0.029392
DeviceProtection_No	0.02127
Dependents_0	0.004193
MonthlyCharges	0.003409
TotalCharges	0.000331

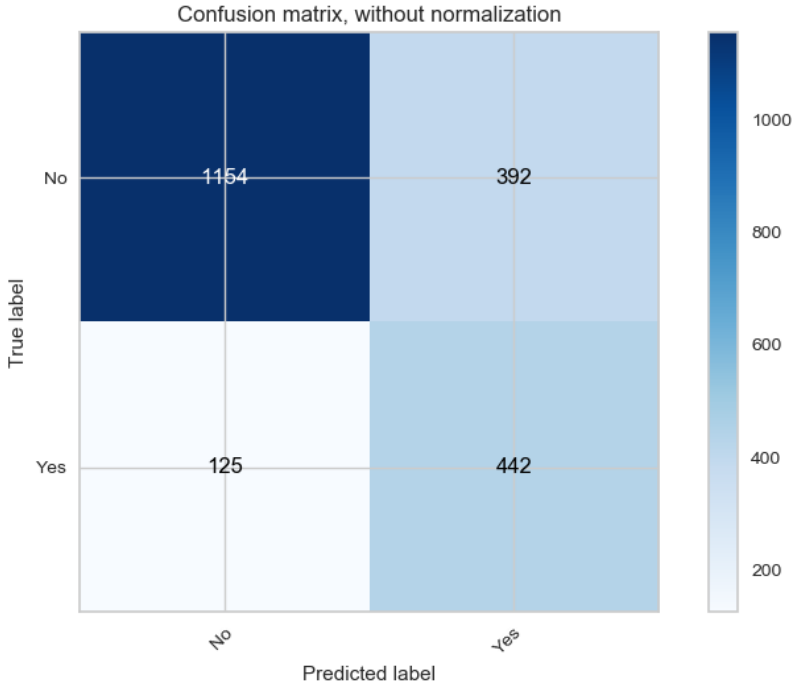
The model and the results

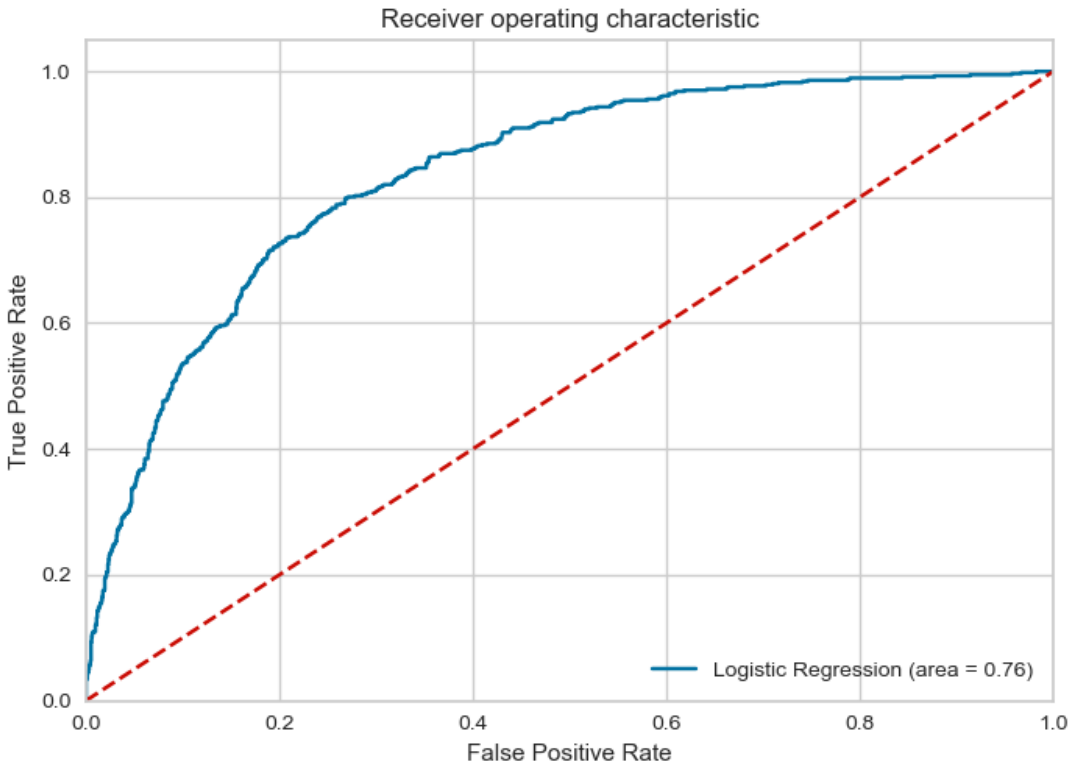
I have implemented two models: logistic regression and Linear SVM.

Logistic regression

The model results look as the following:

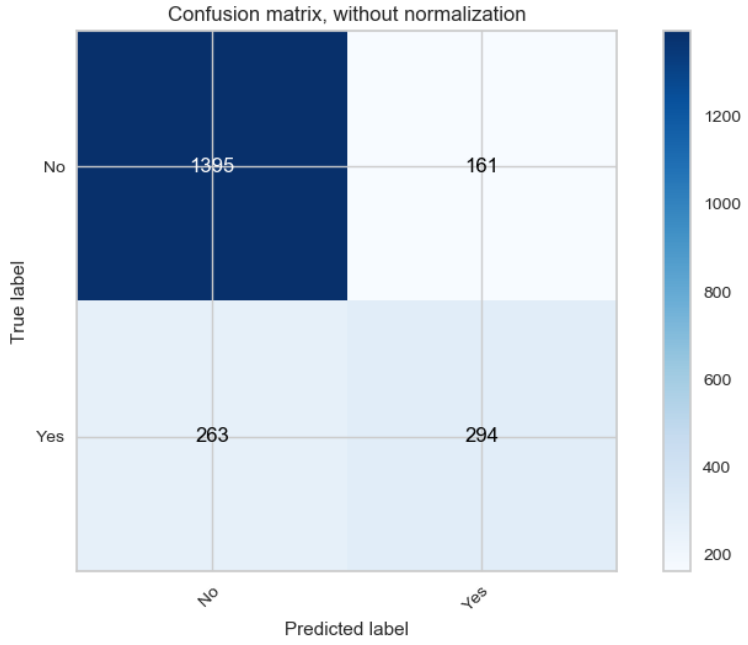

LogisticRegression Classification Report
1
0
0.540
0.767
0.634
0.899
0.760
0.824
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0

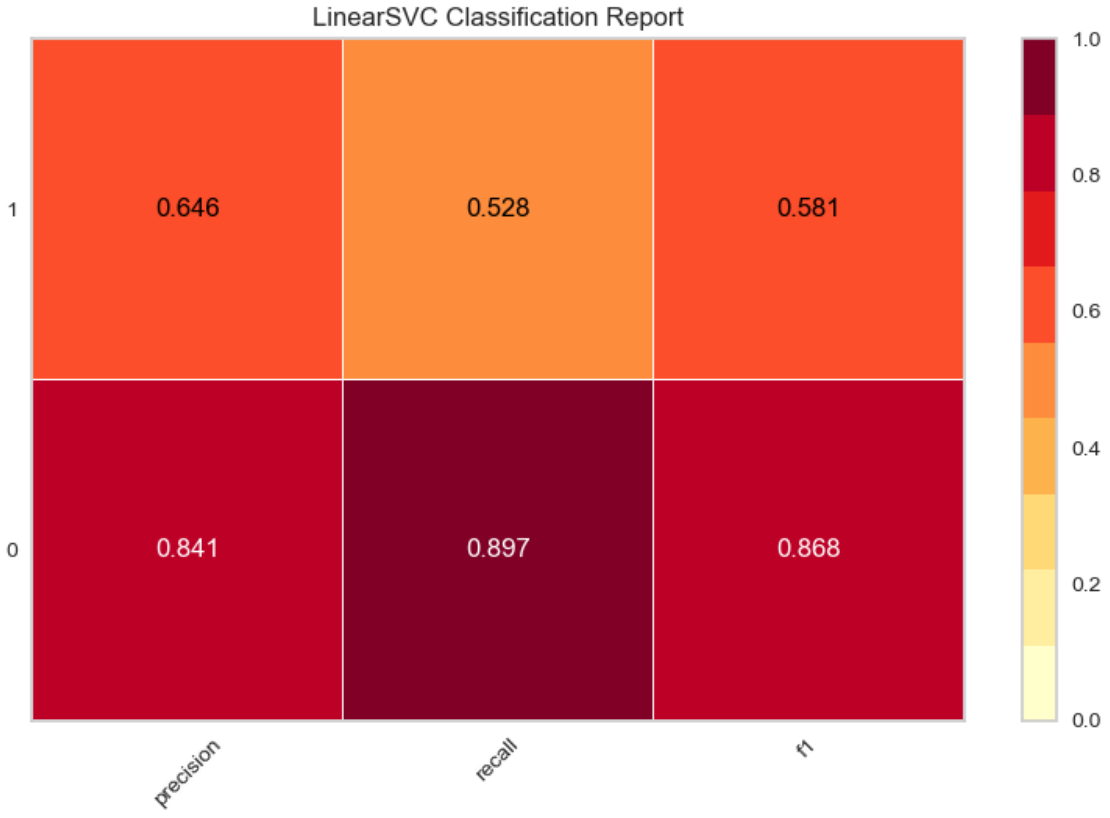

Confusion matrix, without normalization
No
Yes
1154
392
125
442
True label
No
Yes
Predicted label
1000
800
600
400
200



This gives us 76% accuracy, but we must keep in mind that the dataset is imbalanced, and the baseline accuracy is about 73%. This is not a great improvement: the precision of detecting the customers who churned is only 54%, just a bit better than a random chance. But here we have to think of a trade-off: if we want to keep most of the churned customers, the model is quite good – it captures 77% of all churned customers, of course for the cost of capturing 24% of the customers who did not churn.

Linear SVM





In this case the precision is higher (79.93%) and the classification of the customers who did not churn is more correct than in the previous case, but if we have a look on the classification report, the model falsely identified 47% of the customers that churn. I believe that in this case it is better to have more

false positives and higher recall than more false negatives and higher precision. That is why I have picked the Logistic regression as my final model.