Telco Customer Churn Prediction

M V Murali Krishna Nagulla

Introduction

Churn prediction is one of the most popular Big Data use cases in business. It consists of detecting customers who are likely to cancel a subscription to a service.

Churn is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

Overview

- **Goal**: Analyze telecommunication company customer data to predict whether or not a customer is likely to leave the platform (churn)
- Data from 7043 customers (21 features):
  - Churn (Yes or No)
  - Customer account information (tenure, contract, payments, etc.)
  - Demographic Information (Partner, Gender, Age, etc.)
  - Add-on services provided by the platform
- Source: Kaggle.com

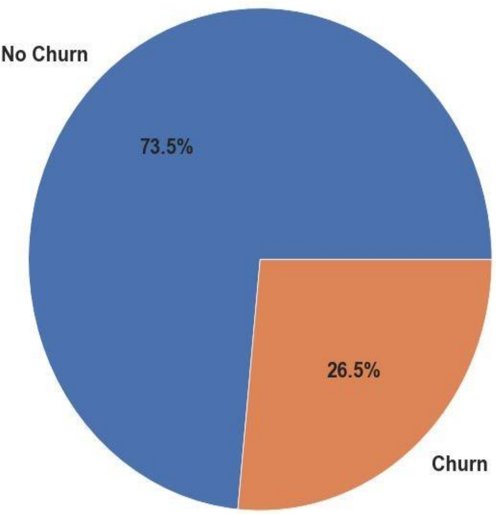
Project Objective

- To predict Customer Churn.
- Highlighting the main variables/factors influencing Customer Churn.
- Use various ML algorithms to build prediction models, evaluate the accuracy and performance of these models.
- Finding out the best model for our business case & providing executive summary.

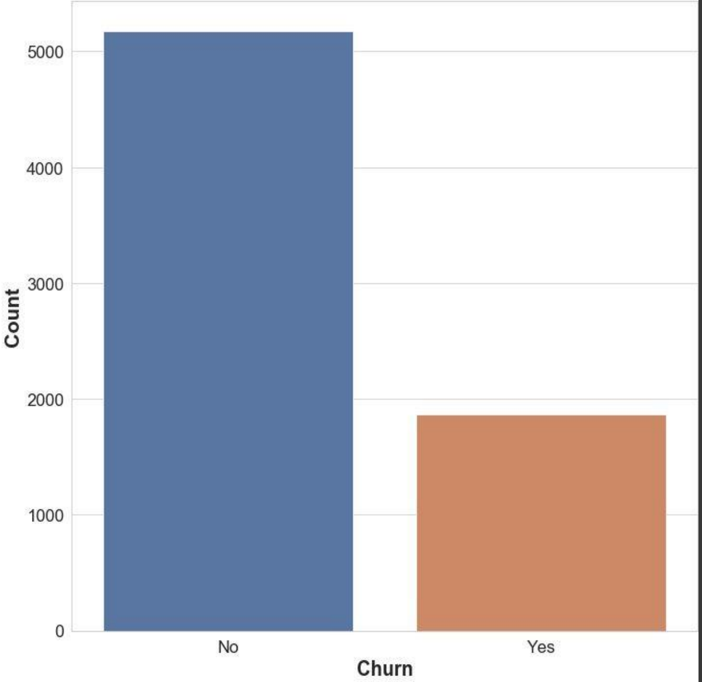
Dataset

- Data
- 20 columns(Removed ID), 7032 entries
- Mostly Categorical Data

Churn Outcome Pie Chart



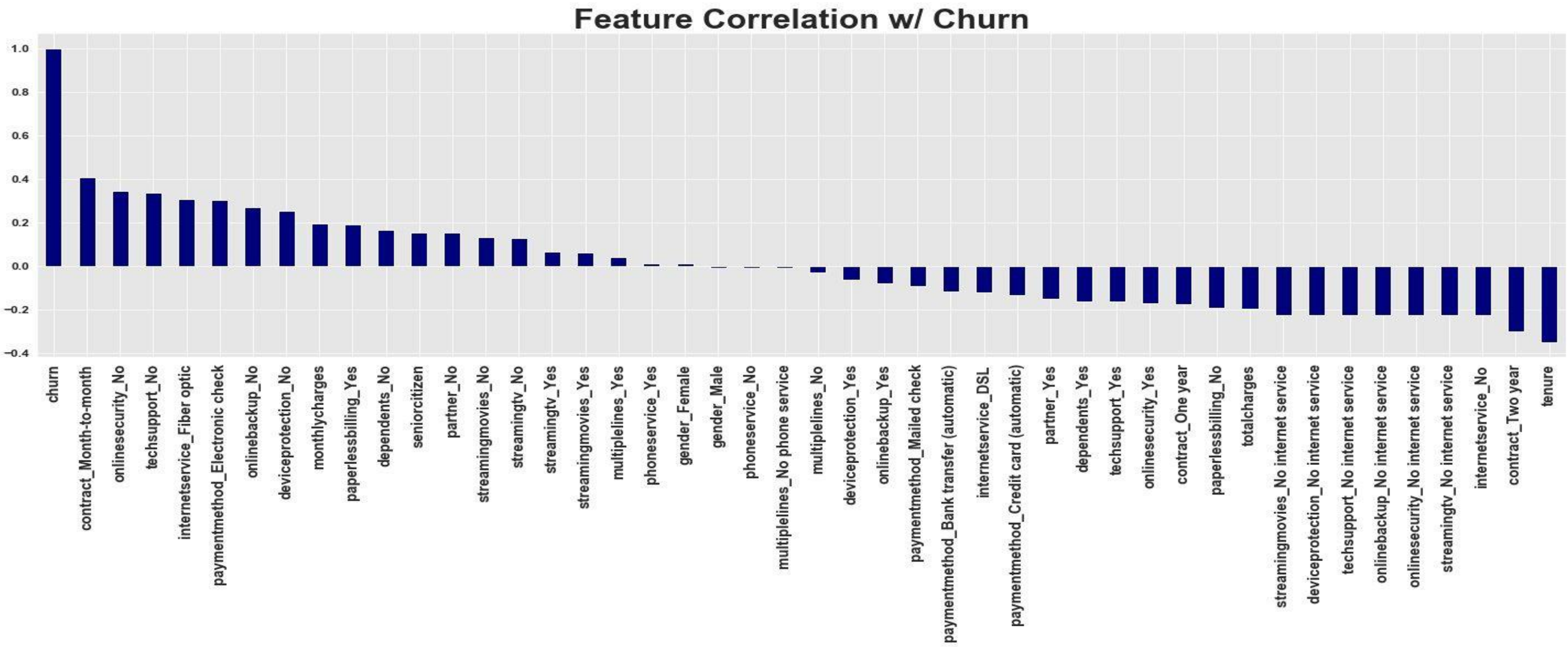
Churn Outcome Distributions




```
[ ]  df.info()

     <class 'pandas.core.frame.DataFrame'>
     Int64Index: 7032 entries, 0 to 7042
     Data columns (total 20 columns):
      #   Column            Non-Null Count  Dtype
     ---  ------            --------------  -----
      0   gender            7032 non-null   object
      1   seniorcitizen     7032 non-null   int64
      2   partner           7032 non-null   object
      3   dependents        7032 non-null   object
      4   tenure            7032 non-null   int64
      5   phoneservice      7032 non-null   object
      6   multiplelines     7032 non-null   object
      7   internetservice   7032 non-null   object
      8   onlinesecurity    7032 non-null   object
      9   onlinebackup      7032 non-null   object
      10  deviceprotection  7032 non-null   object
      11  techsupport       7032 non-null   object
      12  streamingtv       7032 non-null   object
      13  streamingmovies   7032 non-null   object
      14  contract          7032 non-null   object
      15  paperlessbilling  7032 non-null   object
      16  paymentmethod     7032 non-null   object
      17  monthlycharges    7032 non-null   float64
      18  totalcharges      7032 non-null   float64
      19  churn             7032 non-null   object
     dtypes: float64(2), int64(2), object(16)
     memory usage: 1.1+ MB
```

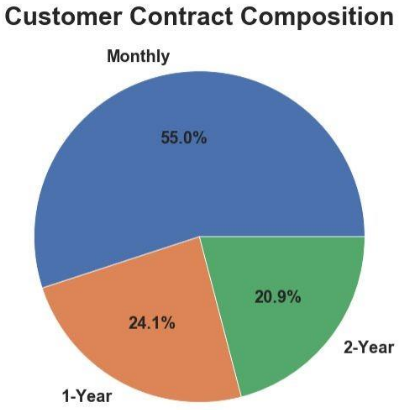
Feature Correlation to Churn

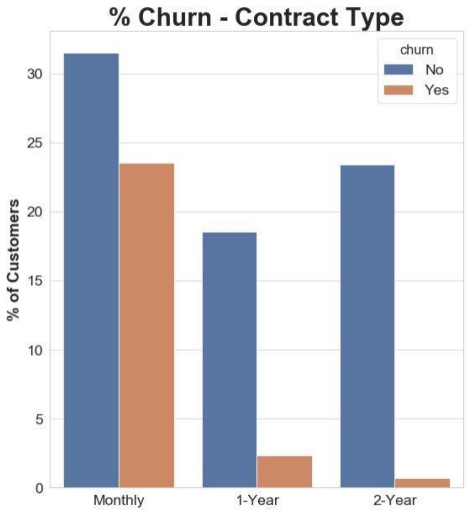


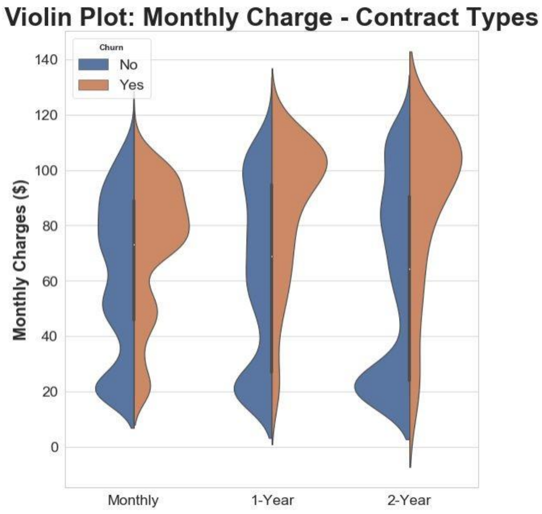
Most Positively Correlated: Monthly Contracts, No Online Security Add-On

Most Negatively Correlated: Tenure, Two-Year Contracts, No Internet Service

Contract Types

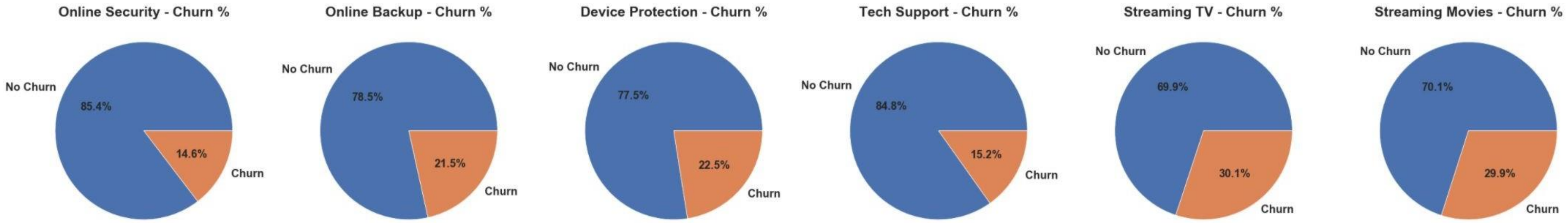






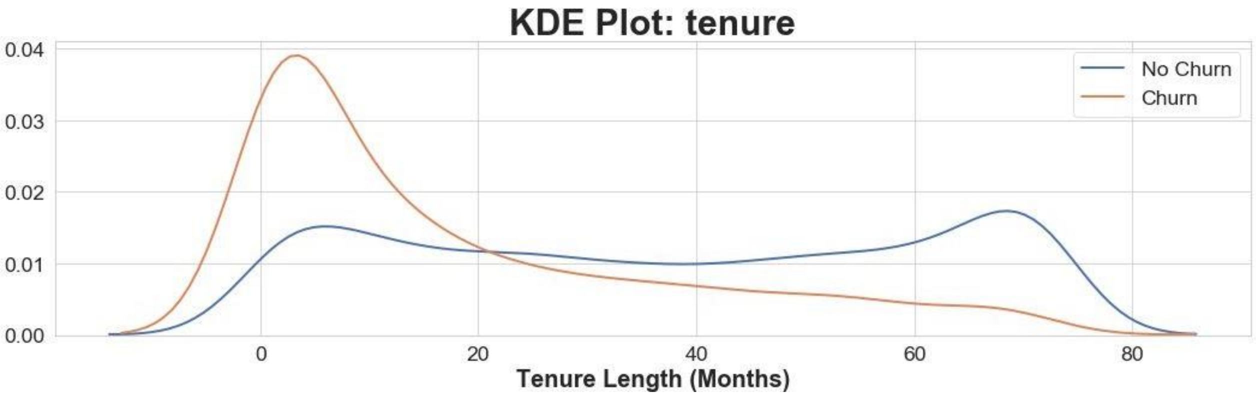
- Significantly more customers will churn when on monthly contracts
- Churn decreases as contract lengths increase
- Customers on monthly contracts are most likely to pay above $60/bill

Add-On Services



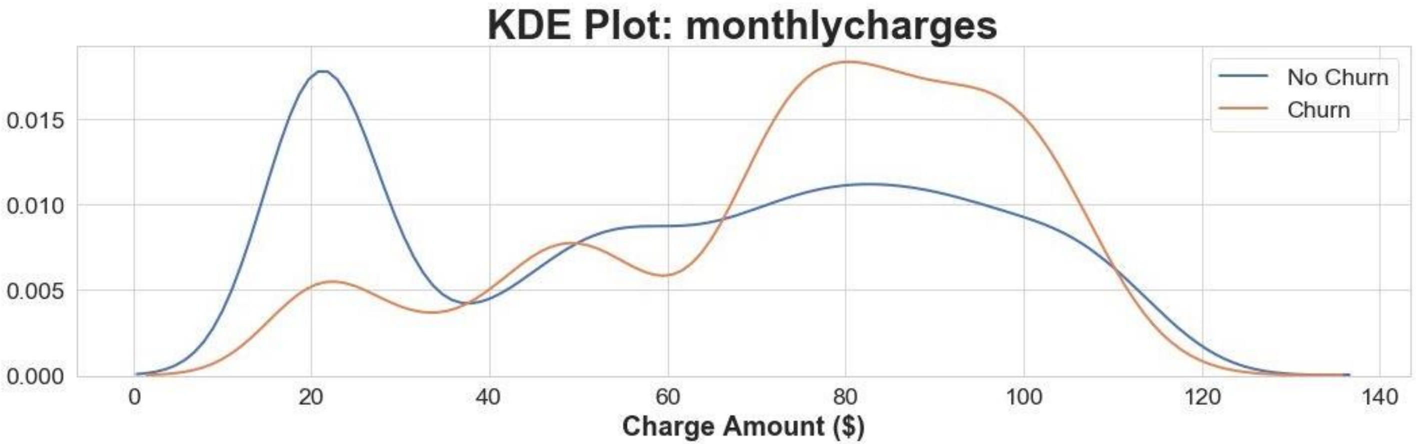
- Customers with online security and/or tech support add-ons will churn the least
- Customers with Streaming services (TV/Movies) will churn the most

Customer Tenure



- Customers are significantly more likely to churn within the first year of tenure on the platform
- As tenure increases, probability of churn decreases

Customer Monthly Charges



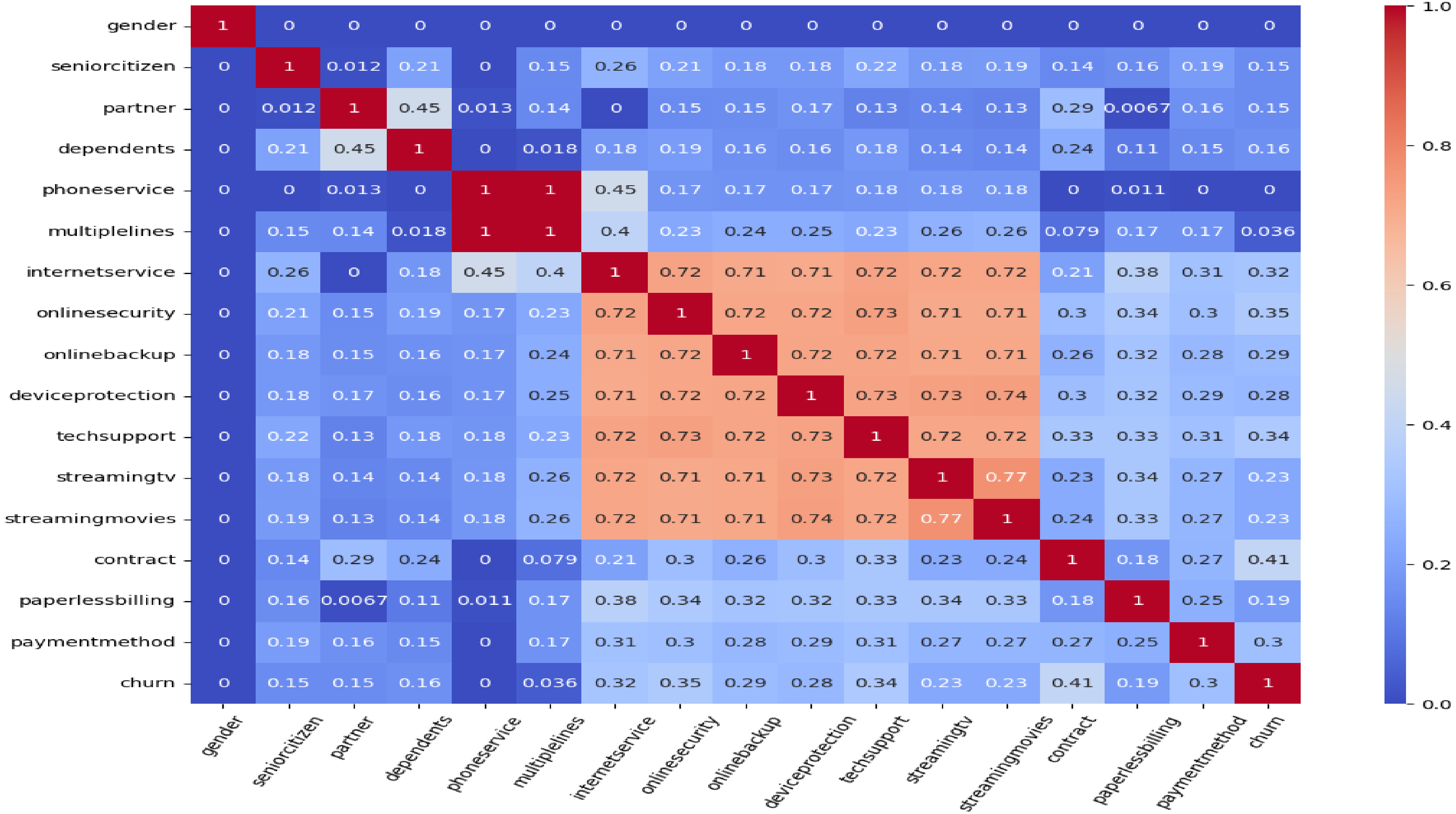
- As monthly charges increase, the probability of customer churn increases
- Customers who churn are most likely to have bills exceeding $60

EDA Conclusions

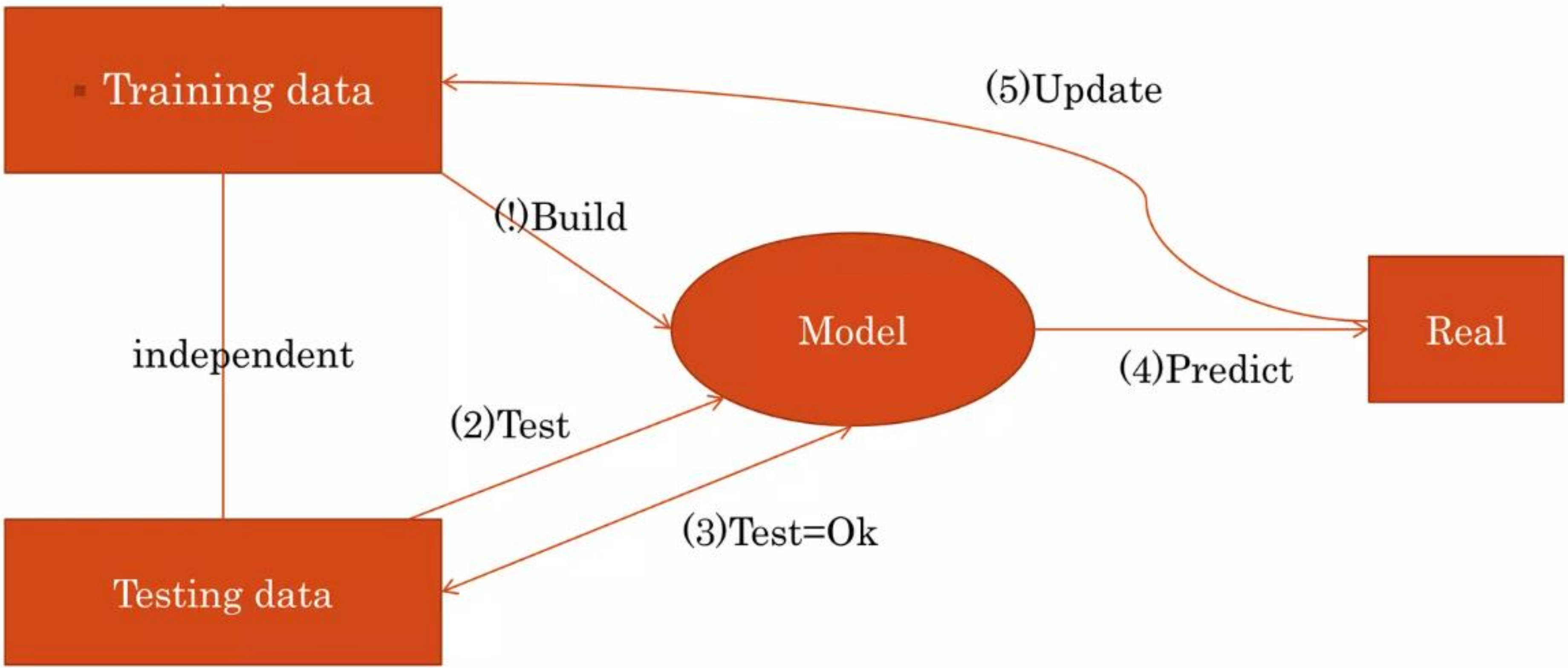
- Customers with monthly contracts are 20% more likely to churn than with annual contracts
- Customers have the highest probability of churning within the first 20 months on the platform
- Tech-Support & Online Security add-ons play a critical role in preventing churn, while streaming add-ons significantly increase likelihood of churn
- Customers are twice as likely to churn when the monthly charge is greater than $60

Correlation

- Inference: There is some correlation between 'phone service' and 'multiple lines' since those who don't have a phone service cannot have multiple lines. So, knowing that a particular customer is not subscribed to phone service we can infer that the customer doesn't have multiple lines.
- Similarly, there is also a correlation between 'internet service' and 'online security', 'online backup', 'device protection', 'streaming tv' and 'streaming movies’.

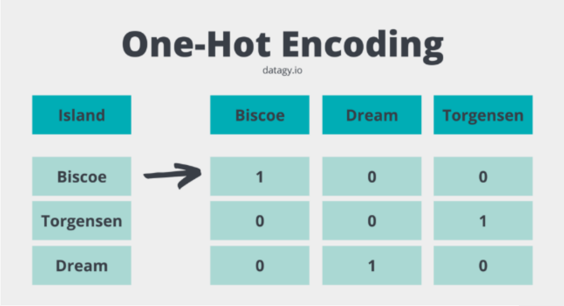

gender
seniorcitizen
partner
dependents
phoneservice
multiplelines
internetservice
onlinesecurity
onlinebackup
deviceprotection
techsupport
streamingtv
streamingmovies
contract
paperlessbilling
paymentmethod
churn
1.0
0.8
0.6
0.4
0.2
0.0

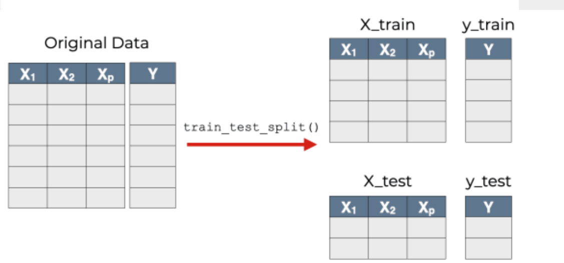
Churn Prediction Model

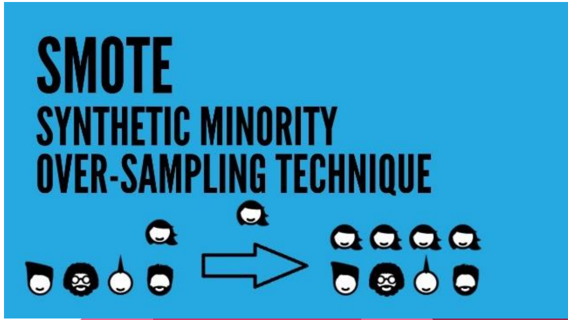


Data Preprocessing

- One-hot-encoding
- Train Test Split
- Oversampling with SMOTE



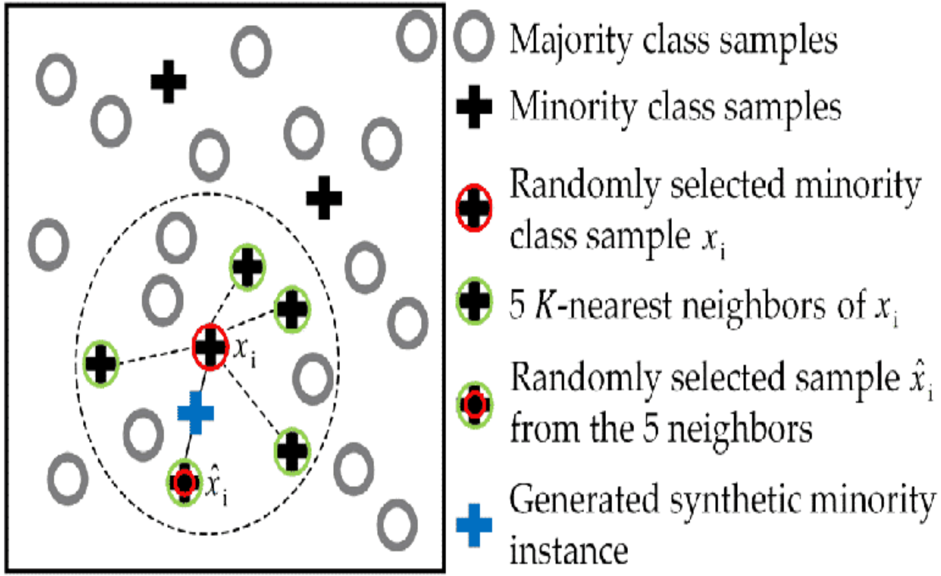




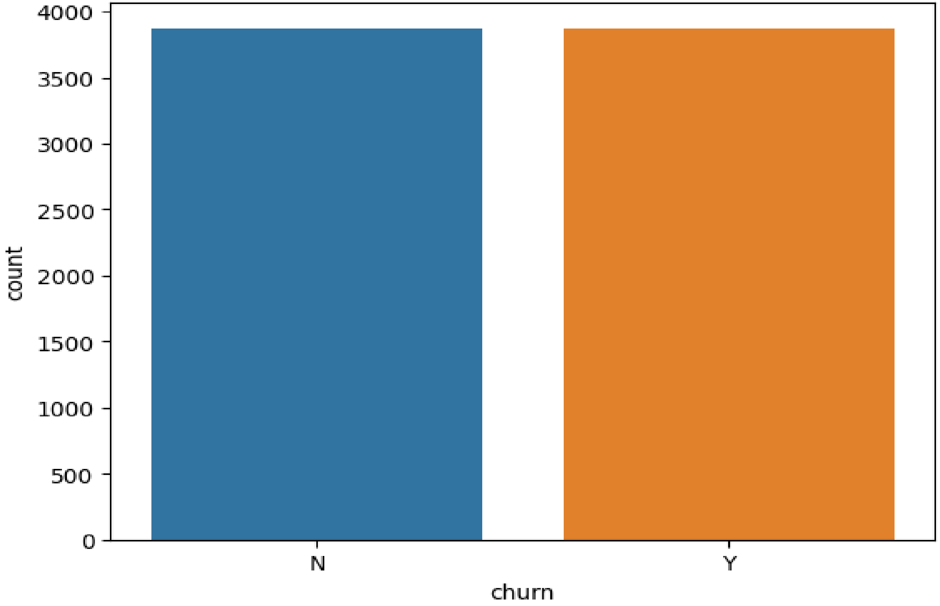
Before Smote

	Train Accuracy	Test Accuracy	Overfitting	ROC Area	Precision	Recall	F1-score	Support
CategoricalNB	0.786	0.735	True	0.821049	0.50137	0.783726	0.611529	467
LogisticRegressionCV	0.812	0.759	True	0.828545	0.53429	0.717345	0.612431	467
KNeighborsClassifier	0.798	0.705	True	0.835032	0.469679	0.845824	0.603976	467
DecisionTreeClassifier	0.791	0.742	True	0.78487	0.513514	0.569593	0.540102	467
BaggingClassifier	0.794	0.716	True	0.833156	0.479695	0.809422	0.60239	467
RandomForestClassifier	0.785	0.714	True	0.828506	0.476923	0.796574	0.596632	467
AdaBoostClassifier	0.821	0.759	True	0.838115	0.533435	0.751606	0.624	467
XGBClassifier	0.849	0.763	True	0.815916	0.543403	0.670236	0.600192	467
Linear SVC	0.809	0.762	True	0.826816	0.53871	0.715203	0.614535	467
RBF SVC	0.83	0.77	True	0.817963	0.557798	0.650964	0.600791	467
CatBoostClassifier	0.856	0.768	True	0.829783	0.554128	0.646681	0.596838	467

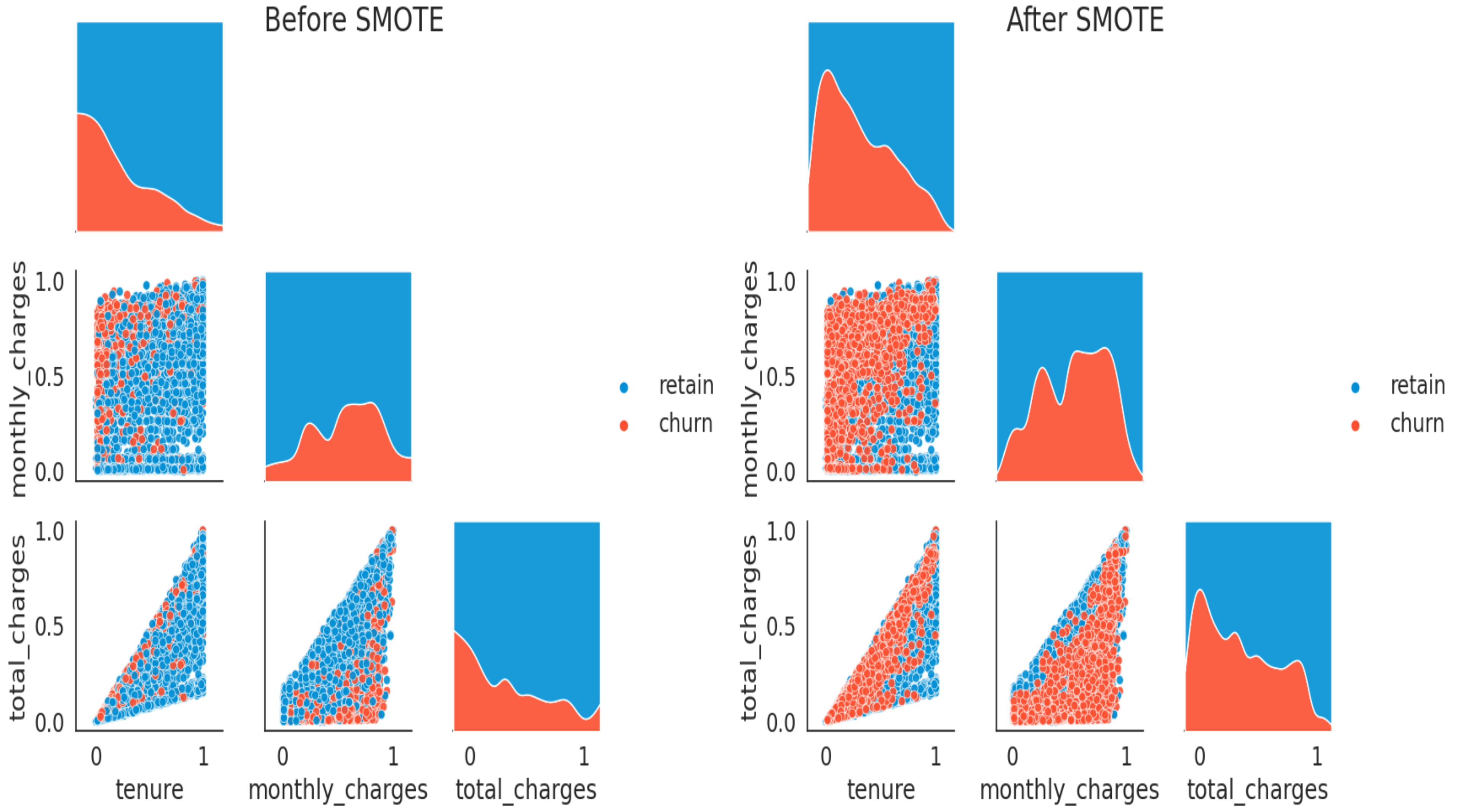
Smote



Majority class samples
Minority class samples
Randomly selected minority class sample x_i
5 K-nearest neighbors of x_i
Randomly selected sample $\hat{x}_i$ from the 5 neighbors
Generated synthetic minority instance
x_i
$\hat{x}_i$



4000
3500
3000
2500
2000
1500
1000
500
0
count
N
Y
churn


Before SMOTE
After SMOTE
monthly_charges
total_charges
1.0
0.5
0.0
0
1
tenure
monthly_charges
total_charges
retain
churn

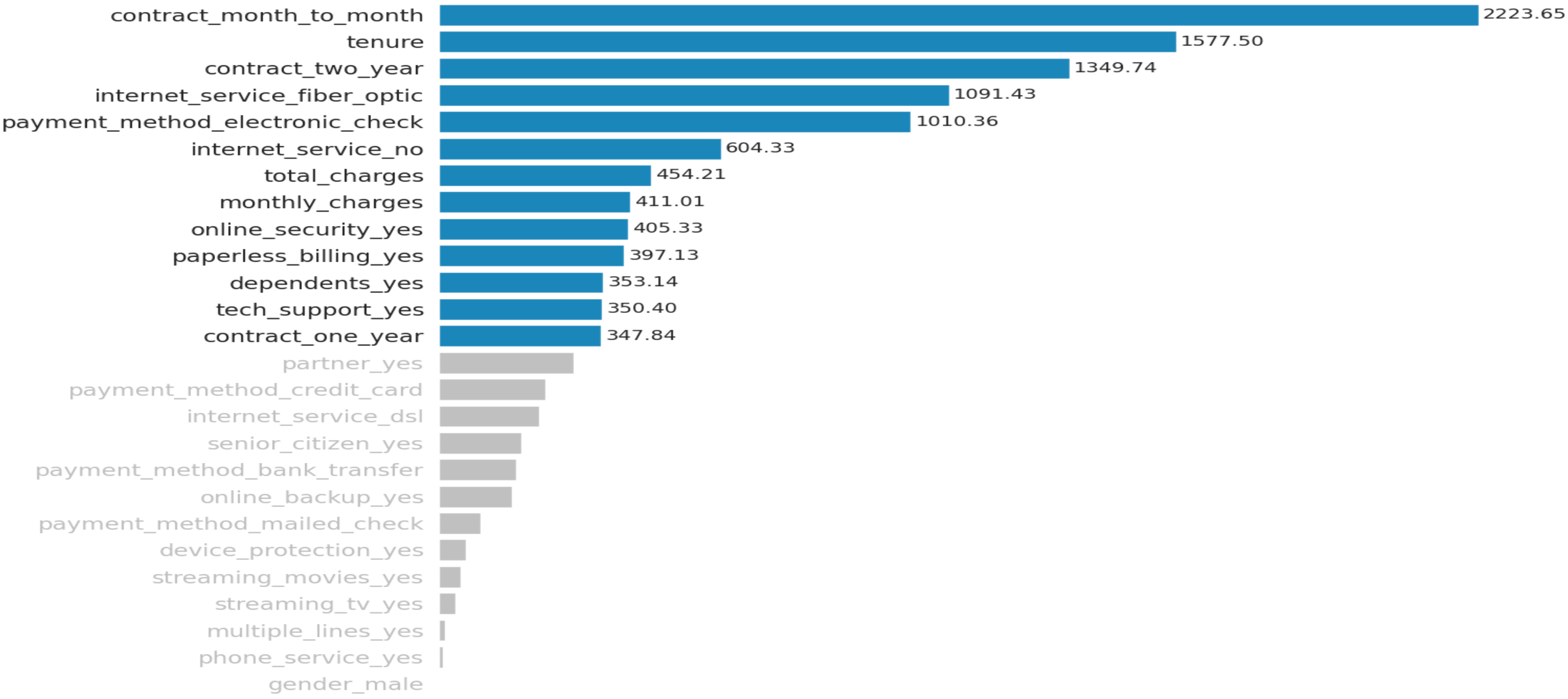
Models

model	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1_score	roc_auc
Logistic Regression	0.746805	0.707463	0.754797	0.714375	0.754797
Ridge Classifier	0.743966	0.706609	0.755140	0.712605	0.755140
KNN	0.696167	0.660272	0.699268	0.661179	0.699268
SVC	0.765736	0.713673	0.747765	0.723961	0.747765
Neural Network	0.758164	0.692219	0.698790	0.695261	0.698790
Decision Tree	0.723616	0.656694	0.671288	0.662195	0.671288
Random Forest	0.773781	0.711190	0.716819	0.713851	0.716819
Gradient Boosting Classifier	0.786559	0.732080	0.758526	0.742016	0.758526
AdaBoost Classifier	0.754851	0.711254	0.755721	0.720050	0.755721
CatBoost Classifier	0.783720	0.723171	0.726430	0.724754	0.726430
Hist Gradient Boosting	0.784193	0.724894	0.734720	0.729374	0.734720
XGBoost	0.778514	0.715981	0.715488	0.715733	0.715488
LightGBM	0.781354	0.720826	0.727665	0.724033	0.727665

Top Performing models

model	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1_score	roc_auc
Gradient Boosting Classifier	0.786559	0.732080	0.758526	0.742016	0.758526
AdaBoost Classifier	0.754851	0.711254	0.755721	0.720050	0.755721
CatBoost Classifier	0.783720	0.723171	0.726430	0.724754	0.726430
Hist Gradient Boosting	0.784193	0.724894	0.734720	0.729374	0.734720
XGBoost	0.778514	0.715981	0.715488	0.715733	0.715488
LightGBM	0.781354	0.720826	0.727665	0.724033	0.727665

Feature Selection



Comparison

I did model performance comparison before and after feature selection. I'll just take the average of each metrics.

	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1_score	roc_auc
original	0.778198	0.721368	0.736425	0.725993	0.736425
filter method	0.772598	0.718596	0.745133	0.726751	0.745133
wrapper method	0.774333	0.719186	0.742710	0.726742	0.742710
embedded method	0.766130	0.712105	0.739781	0.720457	0.739781

Hyperparameter Tuning

model	accuracy	precision	recall	f1_score	roc_auc	accuracy	precision	recall	f1_score	roc_auc
Gradient Boosting Classifier	0.786559	0.581602	0.698752	0.634818	0.758526	0.782773	0.570055	0.739750	0.643910	0.769038
AdaBoost Classifier	0.754851	0.526642	0.757576	0.621345	0.755721	0.764789	0.540404	0.762923	0.632668	0.764194
CatBoost Classifier	0.783720	0.590592	0.604278	0.597357	0.726430	0.770469	0.550398	0.739750	0.631179	0.760661
Hist Gradient Boosting	0.784193	0.587354	0.629234	0.607573	0.734720	0.764316	0.539130	0.773619	0.635432	0.767286
XGBoost	0.778514	0.583184	0.581105	0.582143	0.715488	0.777094	0.563025	0.716578	0.630588	0.757773
LightGBM	0.781354	0.584041	0.613191	0.598261	0.727665	0.760530	0.533414	0.782531	0.634393	0.767554

- After tuning, the accuracy score is mostly decreased.
- But, the recall score has increased dramatically. Therefore, I will use the tuned model for model selection.

Model Selection

- F-beta score to calculate the harmonic mean of accuracy and recall.

- Here I use beta=1, that means the accuracy and recall are considered as equally important. If you more care about recall, you can change β to be higher than 1, and vice versa.
- LightGBM

$$F_\beta = (1+\beta^2)\frac{accuracy * recall}{\beta * accuracy + recall}$$

model	accuracy	recall	fbeta
Gradient Boosting Classifier	0.774255	0.762923	0.768547
AdaBoost Classifier	0.759110	0.784314	0.771506
CatBoost Classifier	0.761477	0.768271	0.764859
Hist Gradient Boosting	0.755797	0.782531	0.768932
XGBoost	0.759110	0.748663	0.753850
LightGBM	0.761477	0.786096	0.773591

Conclusion

Final Model
LightGBM with feature selection using filter method

We should pay more attention to customers who meet the criteria below:
Contract: Month-to-month
Tenure: Short tenure
Internet service: Fiber optic
Payment method: Electronic check
Please, evaluate our service!
Especially for internet service (fiber optic) and payment method (electronic check)

Can we give more benefit to a new customer?
Because the new customer has a high probability to churn