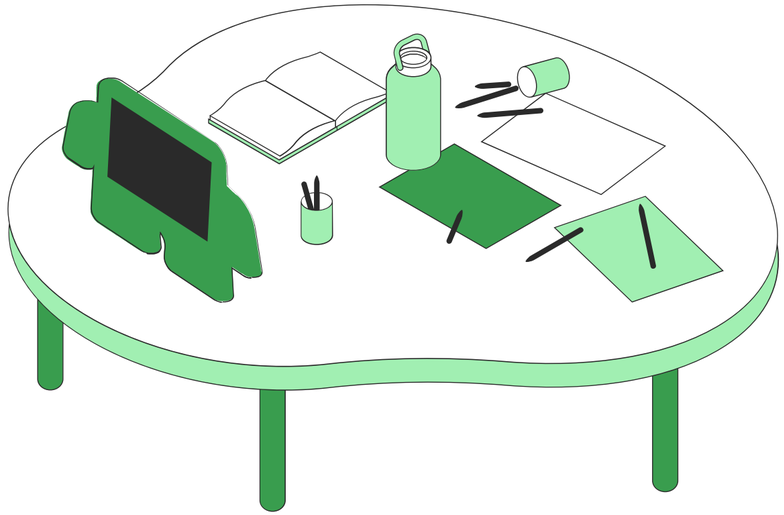
Ma wen zi | PARTNERS

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

By Aymane HANINE

PLAN



Context

Problem

Aproach

Solution

Business impact

Conclusion

What is customer churn?

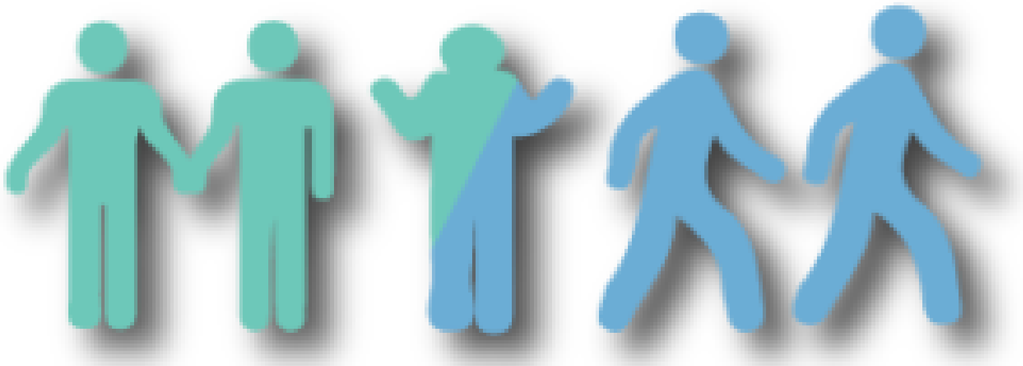
The customer who cease a product or service for a given period is referred as churner.

Customer churn analysis and prediction in ecommerce is an issue now a days because it’s very important to analyze behaviors of various customer to predict which customers are about to leave the subscription.

Problem ?

The main problem is to predict if a customer would leave a subscription/ stop buying or not depending upon the previous data of the customer.

The impact of the churn rate is clear, so we need strategies to reduce it. Predicting churn is a good way to create proactive marketing campaigns targeted at the customers that are about to churn.



A Machine Learning Approach

Thanks to Machine Learning, we can build high performing tool to predict customer churn. The overall scope to build an ML-powered application to forecast customer churn is generic to standardized ML project structure that includes the following steps:

1. **Data Preprocessing**
2. **Data Evaluation**
3. **Model Selection**
4. **Model Evaluation**
5. **Model Improvement**

Data Preprocessing

Columns data types, missing values, unique values...

- The dataset is a 5630 rows × 20 columns table.
- 5 categorical features
- 7 columns containing missing values
- encode categorical data

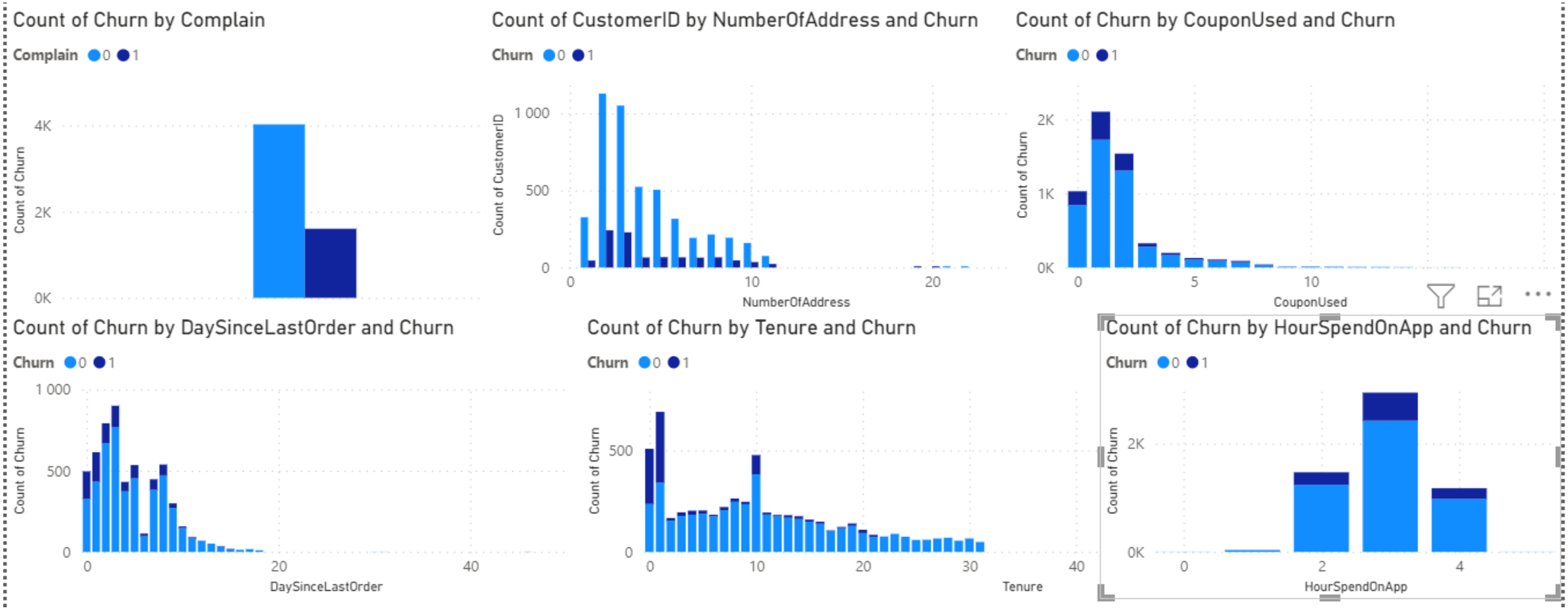
The target column is "Churn"

Churn Value	Count
0	4682
1	948

```
RangeIndex: 5630 entries, 0 to 5629
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   CustomerID                   5630 non-null   int64
 1   Churn                        5630 non-null   int64
 2   Tenure                       5366 non-null   float64
 3   PreferredLoginDevice         5630 non-null   object
 4   CityTier                     5630 non-null   int64
 5   WarehouseToHome              5379 non-null   float64
 6   PreferredPaymentMode         5630 non-null   object
 7   Gender                       5630 non-null   object
 8   HourSpendOnApp               5375 non-null   float64
 9   NumberOfDeviceRegistered     5630 non-null   int64
 10  PreferedOrderCat             5630 non-null   object
 11  SatisfactionScore            5630 non-null   int64
 12  MaritalStatus                5630 non-null   object
 13  NumberOfAddress              5630 non-null   int64
 14  Complain                     5630 non-null   int64
 15  OrderAmountHikeFromlastYear  5365 non-null   float64
 16  CouponUsed                   5374 non-null   float64
 17  OrderCount                   5372 non-null   float64
 18  DaySinceLastOrder            5323 non-null   float64
 19  CashbackAmount               5630 non-null   float64
dtypes: float64(8), int64(7), object(5)
memory usage: 879.8+ KB
```

Data Evaluation

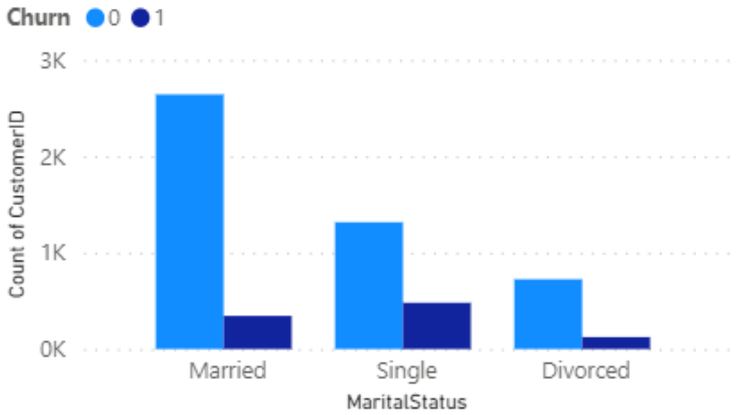
Plot histogram of numeric Columns



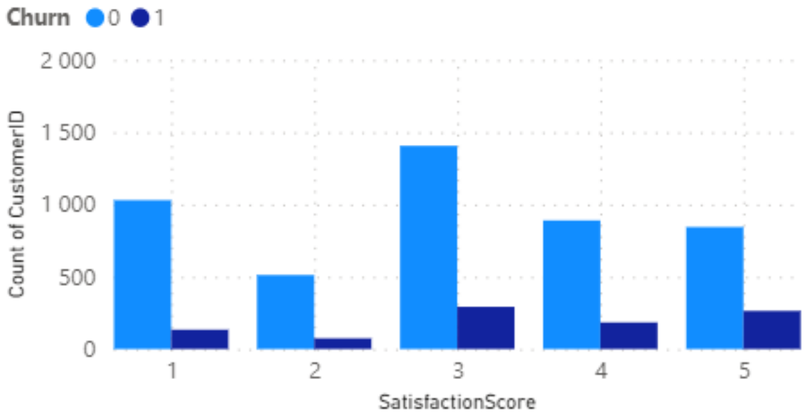
Data Evaluation

Analyze the distribution of categorical variables

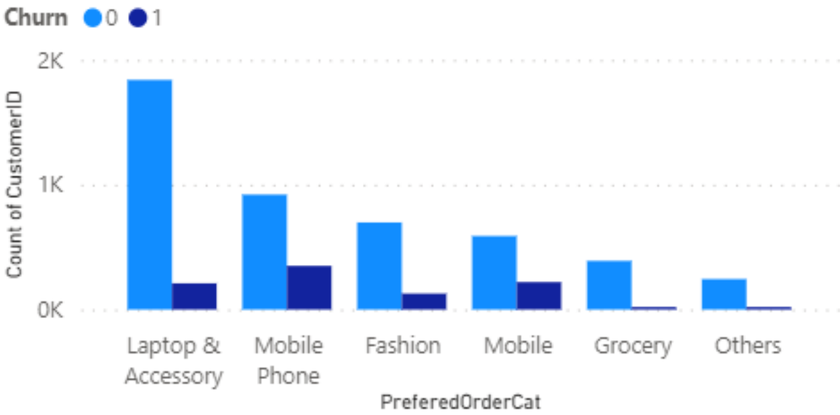
Count of CustomerID by MaritalStatus and Churn



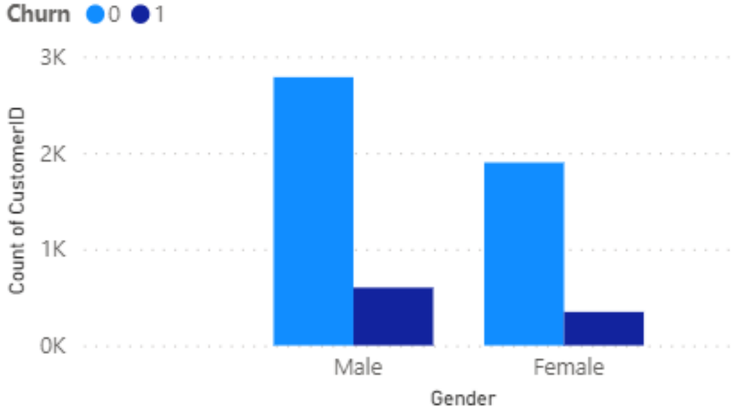
Count of CustomerID by SatisfactionScore and Churn



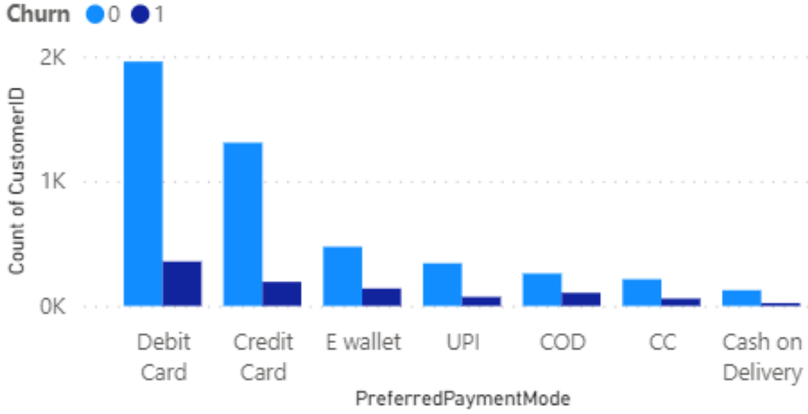
Count of CustomerID by PreferedOrderCat and Churn



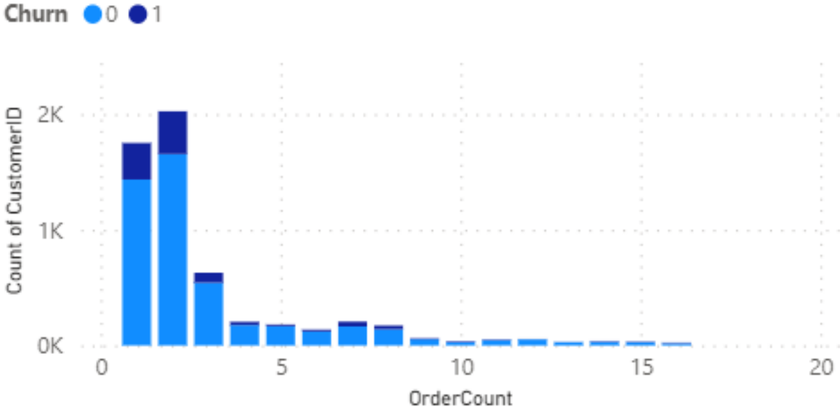
Count of CustomerID by Gender and Churn



Count of CustomerID by PreferredPaymentMode and Chu



Count of CustomerID by OrderCount and Churn



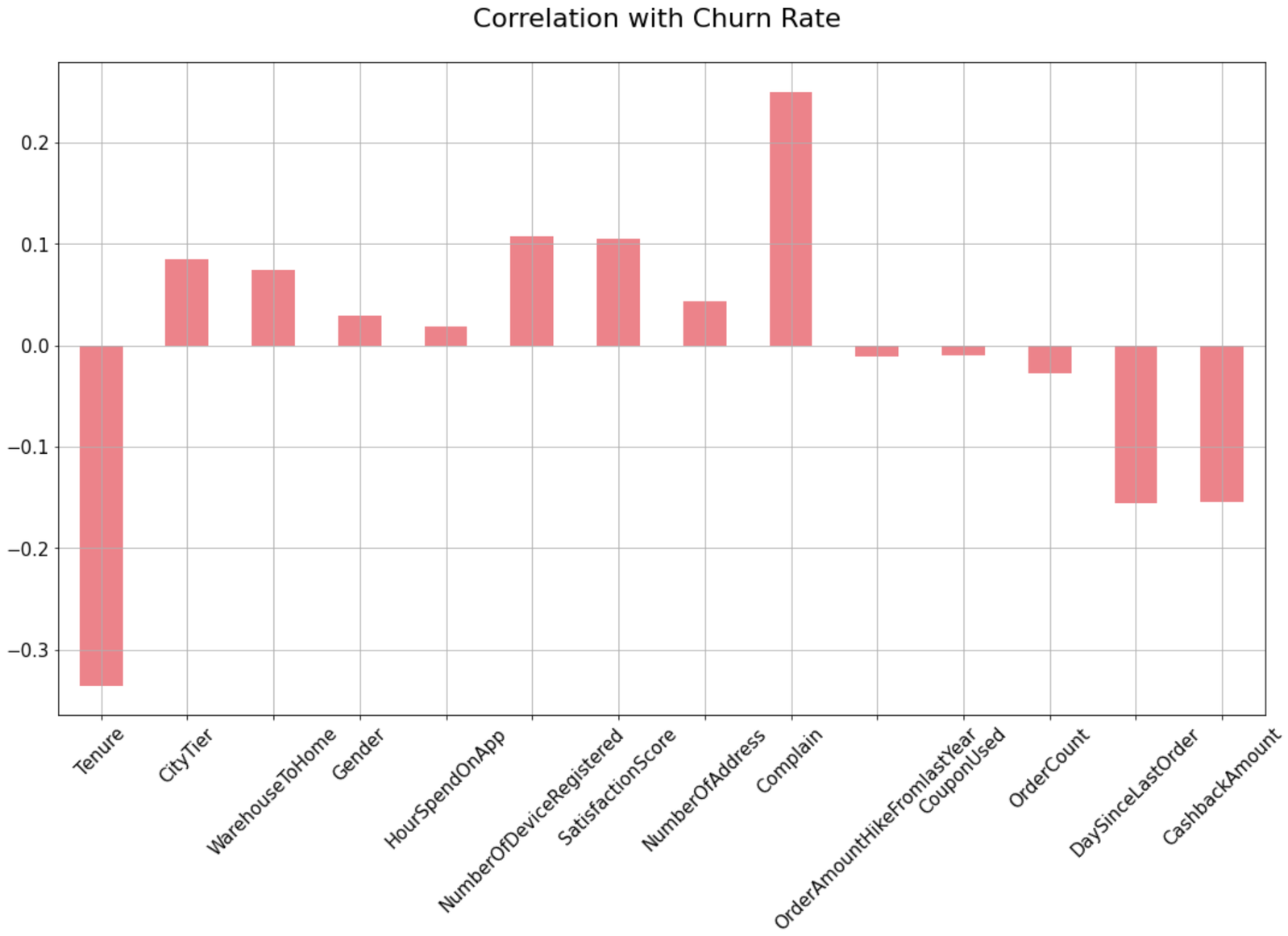
Data Evaluation

Plot positive & negative correlations

Correlation matrix helps us to discover the bivariate relationship between independent variables in a dataset.

In this case, we can see a good positive correlatio in "Complain". there is a high probability that a customer churn if he has already complained.

Also, the higher the tenue, day of last order and cashback are the less chance that a customer will churn.



Model Selection

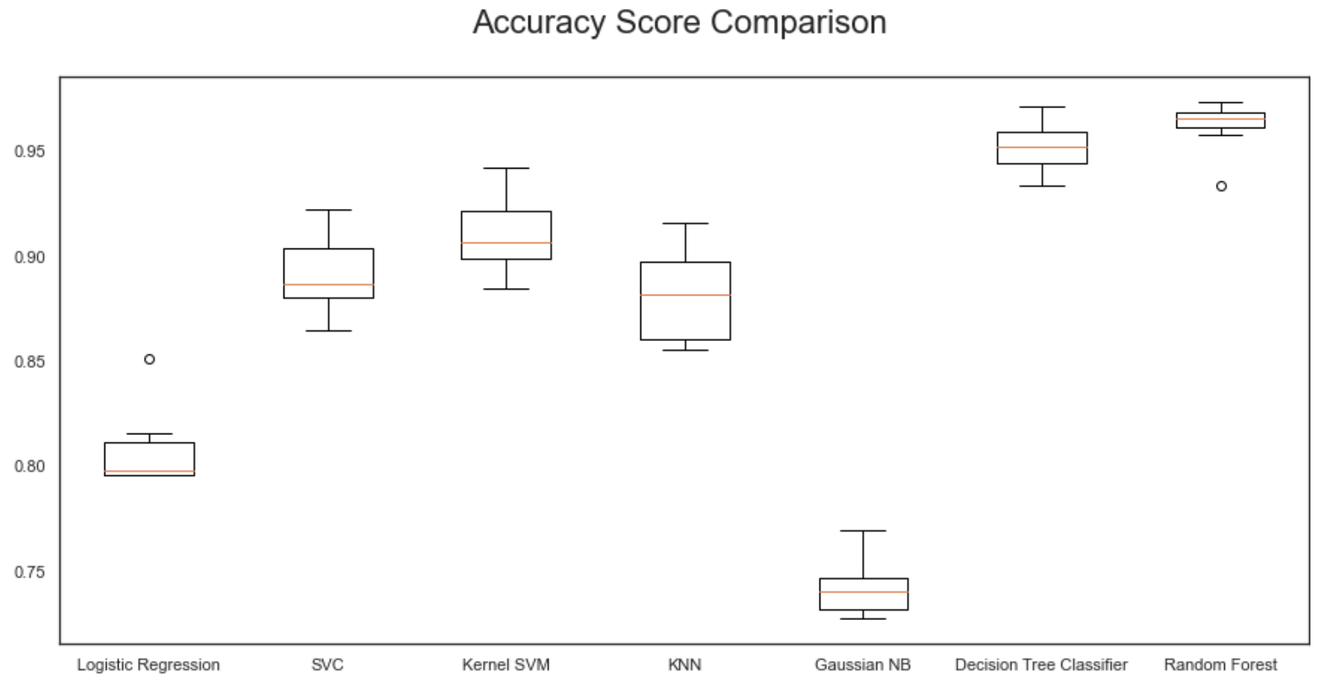
Compare Baseline Classification Algorithms

I modeled each classification algorithm over the training dataset and evaluate their accuracy and standard deviation scores.

To select the best model, I have used the Accurancy of the model. which is

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

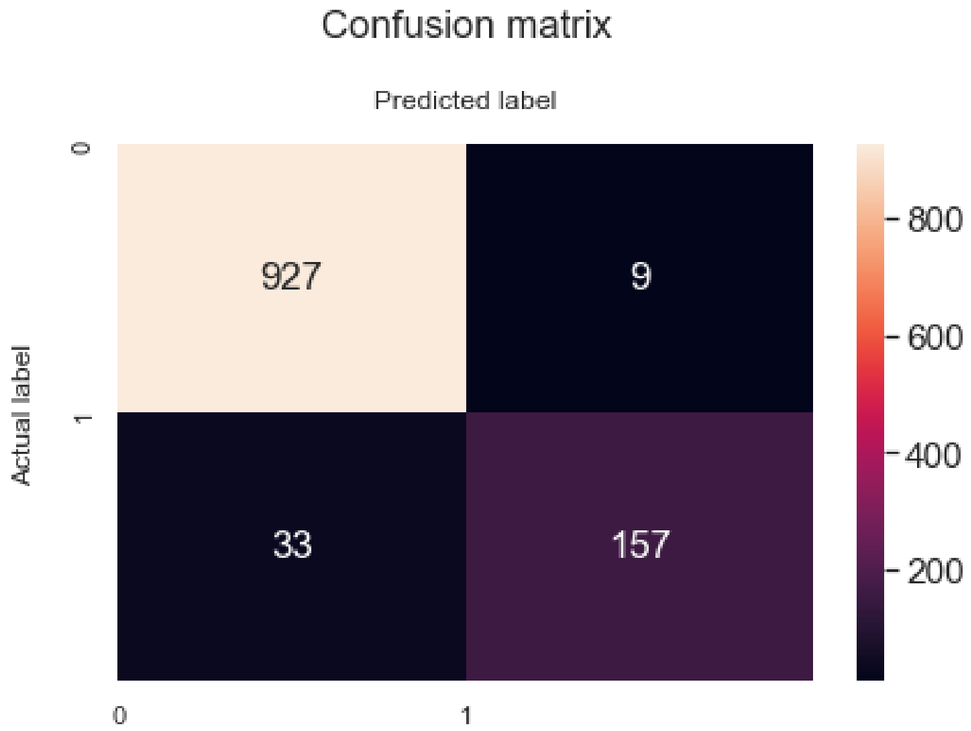
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



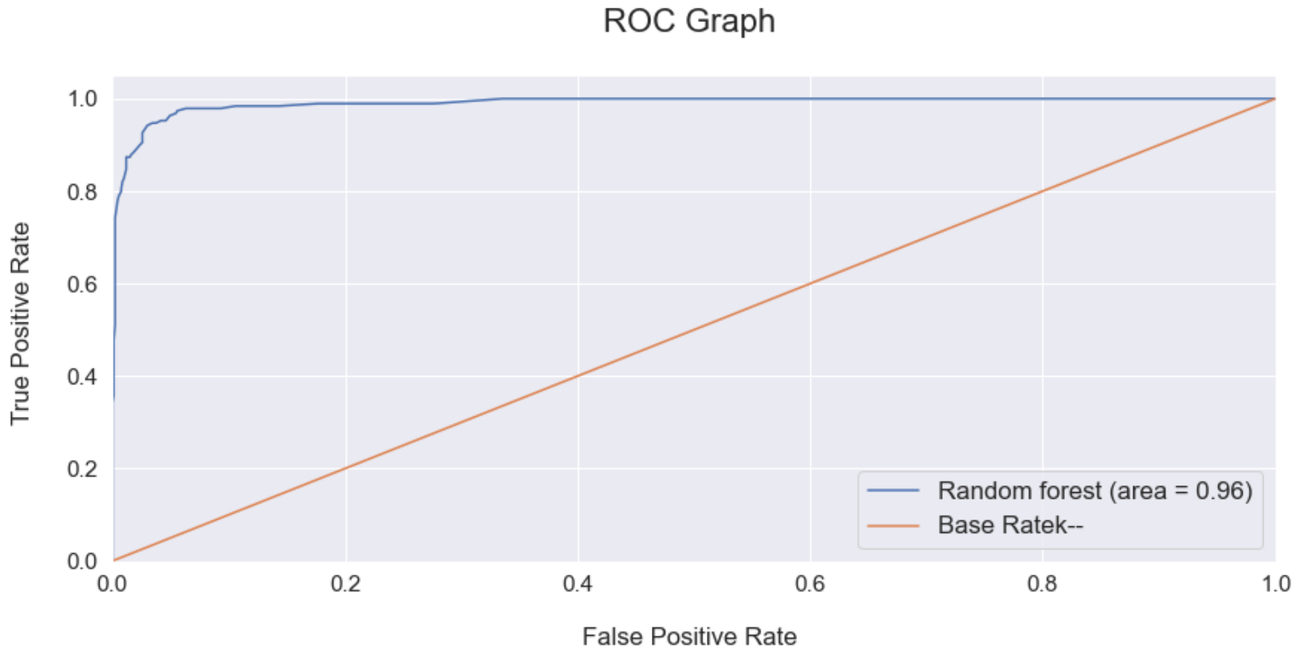
	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
6	Random Forest	98.88	0.54	96.27	1.08
2	Kernel SVM	92.61	1.24	91.01	1.65
5	Decision Tree Classifier	91.83	2.30	95.18	1.15
3	KNN	89.95	1.72	88.10	2.02
0	Logistic Regression	89.22	1.74	80.66	1.65
1	SVC	89.12	1.97	89.14	1.71
4	Gaussian NB	77.63	2.41	74.22	1.33

Model Evaluation

To evaluate the model, I have run a 'K- fold Cross-Validation' technique that primarily helps us to fix the variance. Variance problem occurs when we get good accuracy while running the model on a training set and a test set but then the accuracy looks different when the model is run on another test set.



Random Forest Classifier Accuracy: 0.96 (+/- 0.02)



Model Improvement

Model improvement basically involves choosing the hyperparameters, a set of configurable values external to a model that cannot be determined by the data, for the machine learning model that we have come up with.

To do so I used a grid search lethod as shown in the code.

In this case, the hyperparameters tuning didn't increase the accuracy of the model.

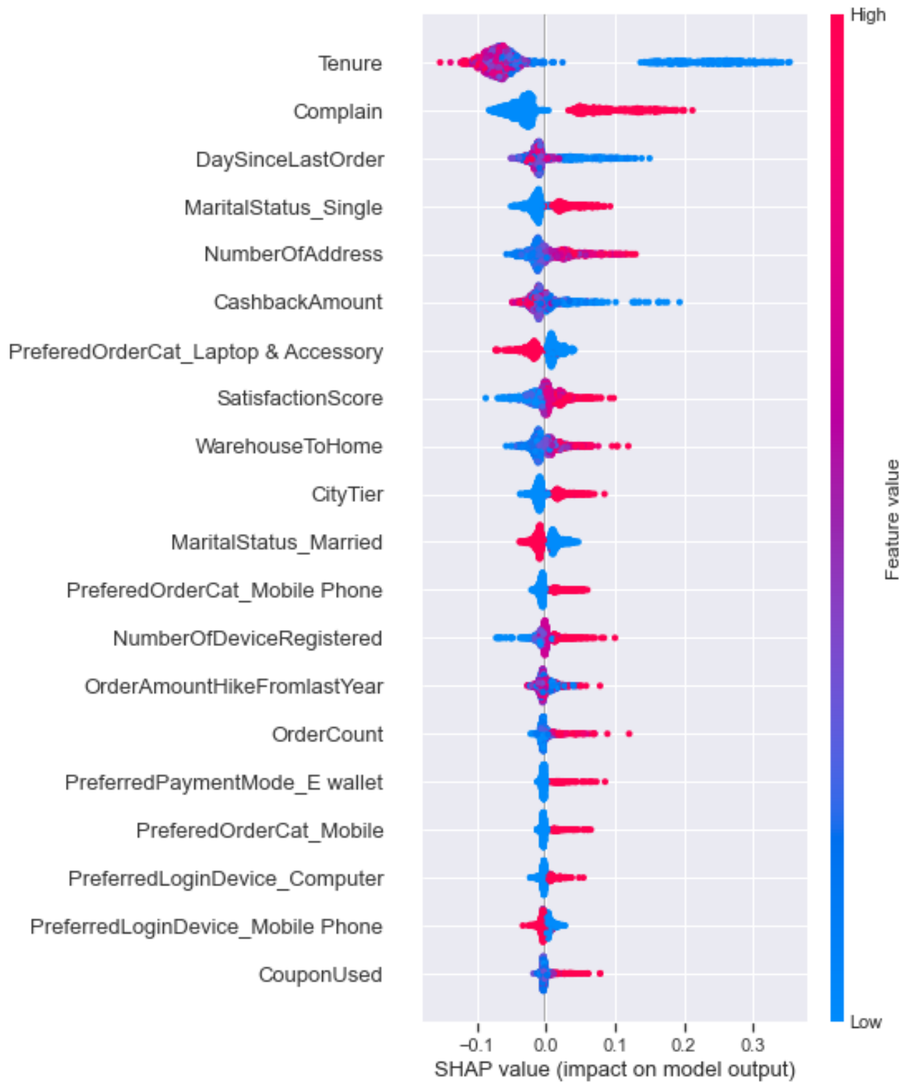
```
from sklearn.model_selection import GridSearchCV
# Create the parameter grid based on the results of random search
param_grid = {
    'bootstrap': [True],
    'max_depth': [80, 90, 100, 110, None],
    'max_features': [2, 3, "sqrt"],
    'min_samples_leaf': [1, 3, 4, 5],
    'min_samples_split': [2, 8, 10, 12],
    'n_estimators': [100, 200, 300, 1000]
}
# Create a based model
rf = RandomForestClassifier()
# Instantiate the grid search model
grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                           cv = 3, n_jobs = -1, verbose = 2)
```

```
{'bootstrap': True,
 'max_depth': 110,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 1000}
```

→ Best Model Parametes

Model Explanability

SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.



Business Impact

- Avoid the loss of revenue that results from a customer abandoning the business
- Know which marketing actions will have the greatest retention impact on each particular customer

Conclusion & perspectives

Customer churn prediction is crucial to the long-term financial stability of a company.

In this work, we have discovered a few machine learning algorithms to predict the churn. We can optimize futher more these algorithms and test other algorithms of Deep Learning.

Thank you for your attention