Churn Prediction

Riccardo Possieri
sba23439@student.cct.ie

CCT Dublin, Ireland

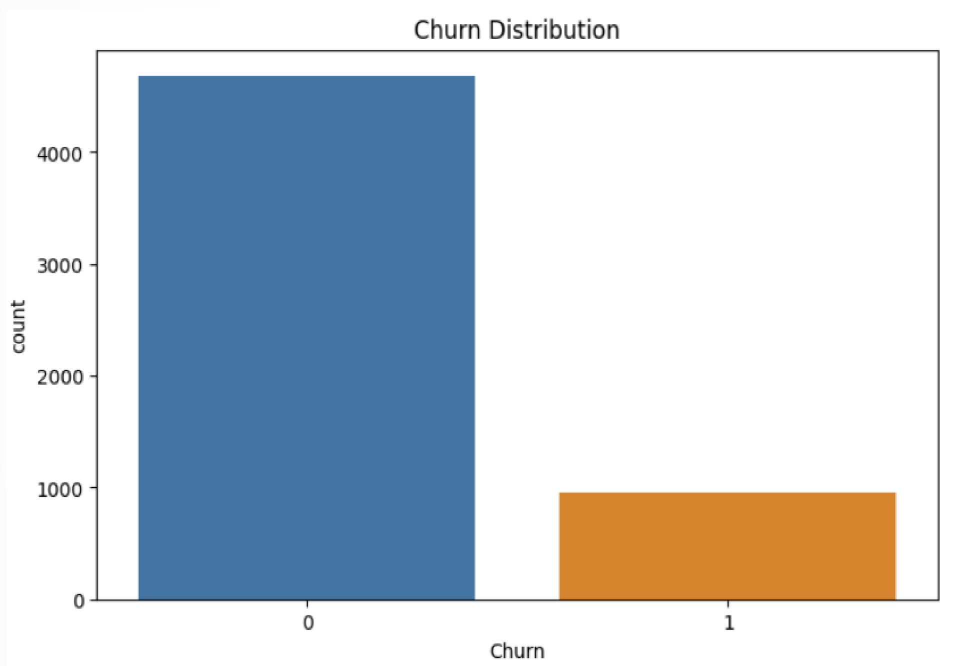
May,2024

Abstract

Churn is a measurement of the percentage of all those accounts that delete or cancel or choose not to renew their subscriptions. Retaining an existing customer is so much cheaper than acquiring a new one. For this reason, companies started to proactively identify customers at risk of churning and implement strategies to retain them. The 'E Commerce Dataset' has been helpful to understand and give us a comprehensive exploration into customer churn within a company.

Problem Formulation & Descriptive Analysis

Our main goal in this project is to create a machine learning model that could predict and help reduce customer churn for the e-commerce company. To achieve this goal, we focus on a key variable called 'Churn.' So, we do 'Churn' feature our target variable which we can see from the plot that it is very imbalanced. The dataset we are going to use has different features and our goal is to determine which ones are the most important to predict the churn. Regarding the dataset: it has 5630 observations and 20 features, 15 numerical and 5 categorical features.



EDA

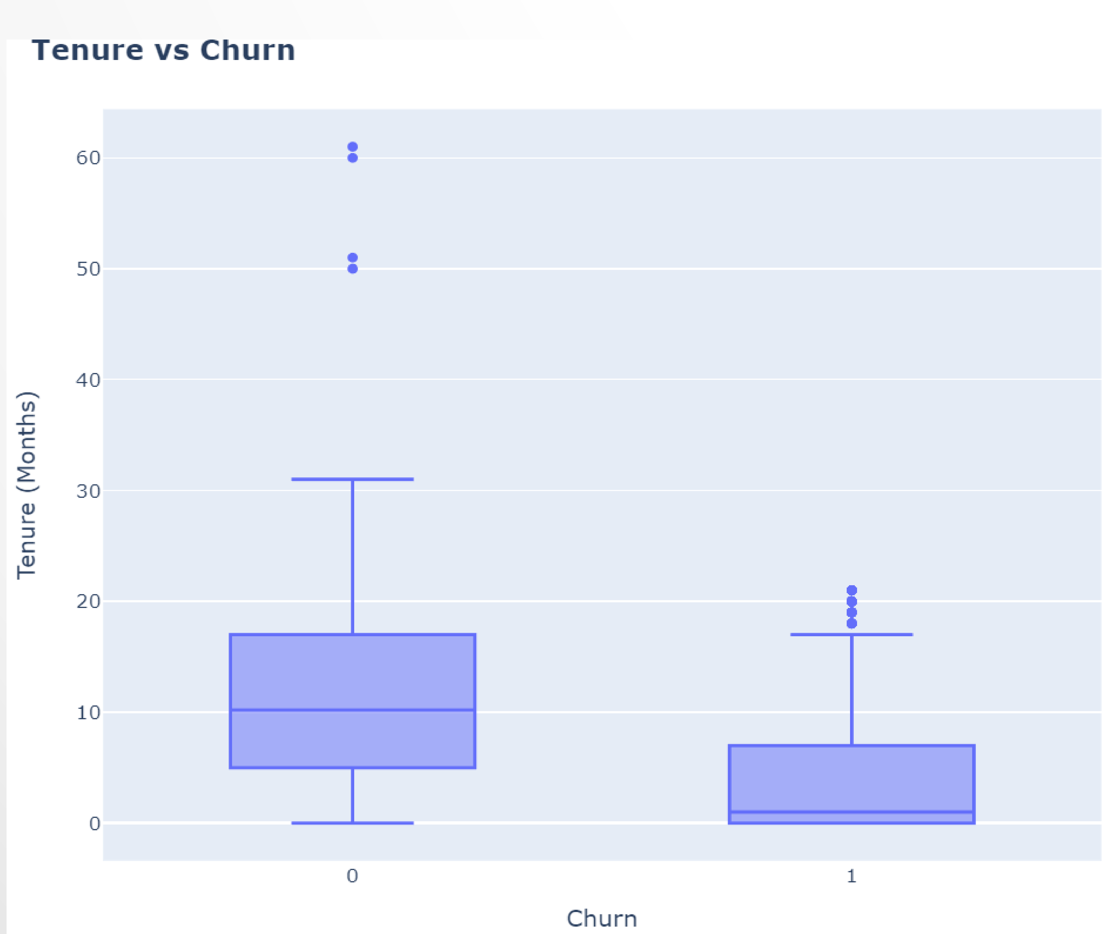
- Is Tenure vs Churn an important relationship to see?
- Is there any realtionship between Churn and Gender?
- Which CityTier has the highest OrderCount?

Gender vs % Churn

Female 15.49%

Male 17.730%

CityTier	OrderCount
1	10835
2	627
3	5471



Data Pre-processing

Point-biserial correlation

$$r_{pb} = \frac{[\overline{M_1} - \overline{M_0}]}{\sigma_y}\sqrt{\frac{n_1 * n_0}{n^2}}$$

Where:

M_1 is the mean of churned customers.

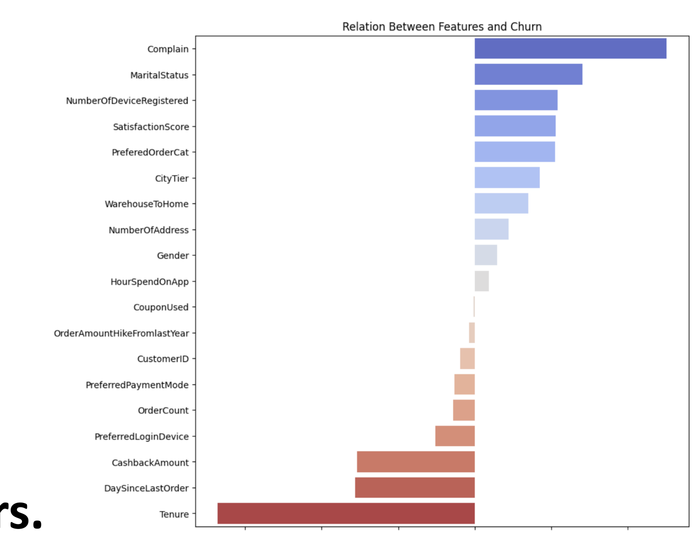
M_0 is the mean of no churned customers.

σ_y is the standard deviation of churn.

n_1 is the number of observations of churned customers.

n_0 is the number of observations of no churned customers.

n is the total number of observations of churn.

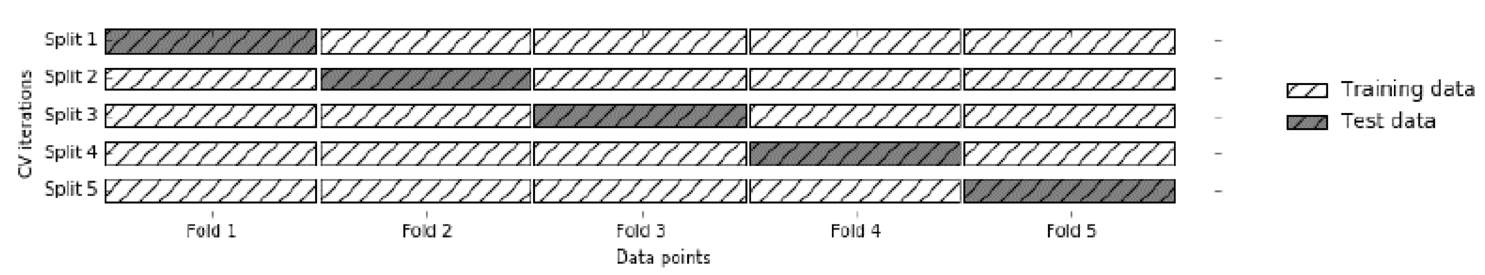


Complain: rpb=0.25

Tenure: rpb=−0.33

Hyperparameter Tuning and Cross-Validation

70% training data, and 30% testing data. Subsequently, we subdivided our training data into 10 splits and 10 folds, using the parameter cv=10



Introduction to Machine Learning with Python (n.d.)

Machine Learning Models

- For logistic regression we used parameters_logreg = {'C': [0.001, 0.01, 0.1, 1, 10, 100]} ,

and we obtained Best Parameter: {'C': 0.1}

- For random forest we used parameters_rf = {'n_estimators': [50, 100, 150, 200],'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4] ,

and we obtained Best Parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

- For support vector machine we used parameters_svm = {'C': [0.1, 1, 10], 'gamma': [0.01, 0.1, 1], 'kernel': ['linear', 'rbf']},

and we obtained Best Parameters: {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}

Initial Results

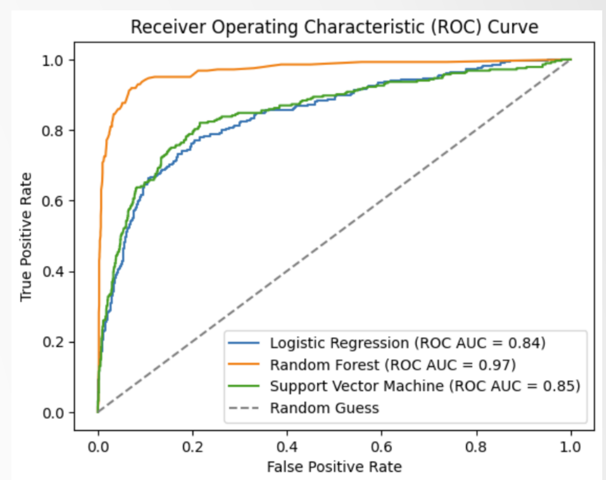
AUC-ROC plots a probability curve about the sensitivity, or TPR against FPR, or $(1 - Specificity)$.

TPR, or $Sensitivity$ or Recall is the formulation of:

$$\frac{TP}{TP + FN}$$

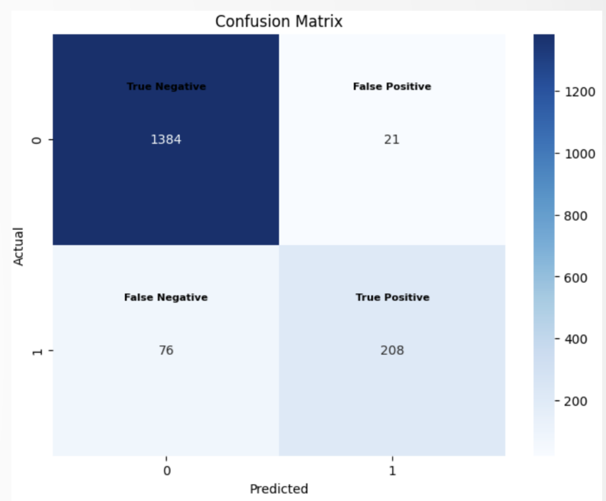
TNR or $Specificity$ is the formulation of:

$$\frac{TN}{TN + FP}$$



And finally, FPR is the formulation of:

$$\frac{FP}{TN + FP} = 1 - Specificity$$

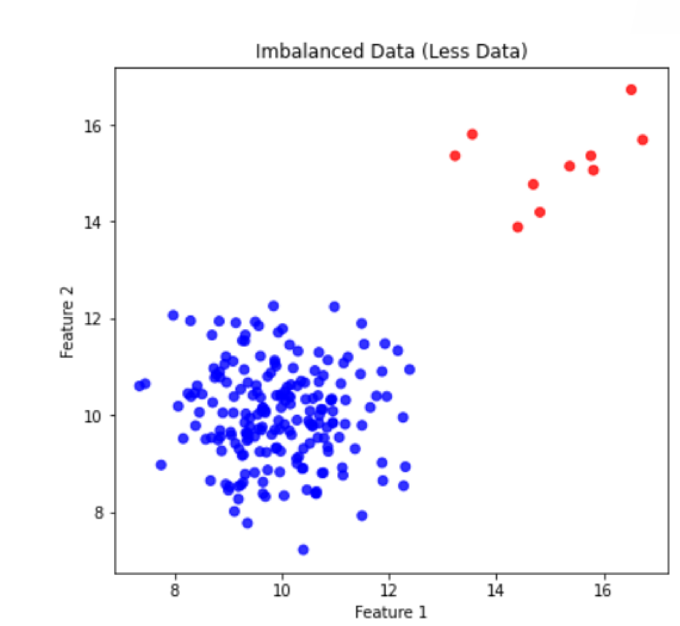


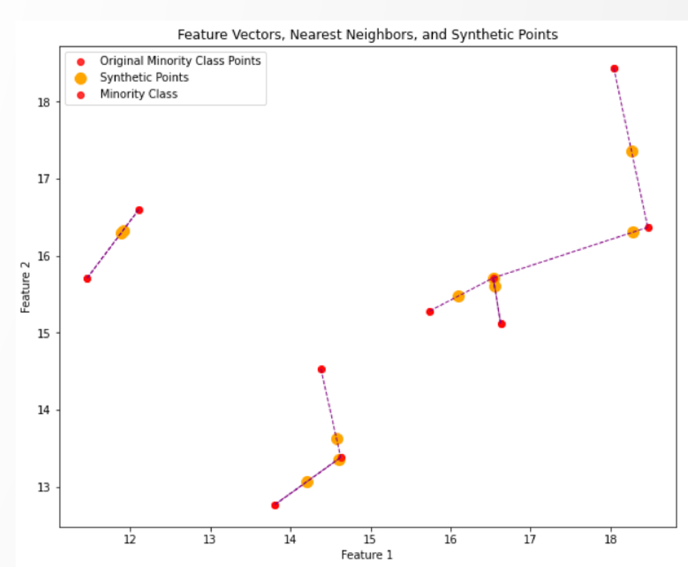
Logistic Regression ROC AUC: 0.8430

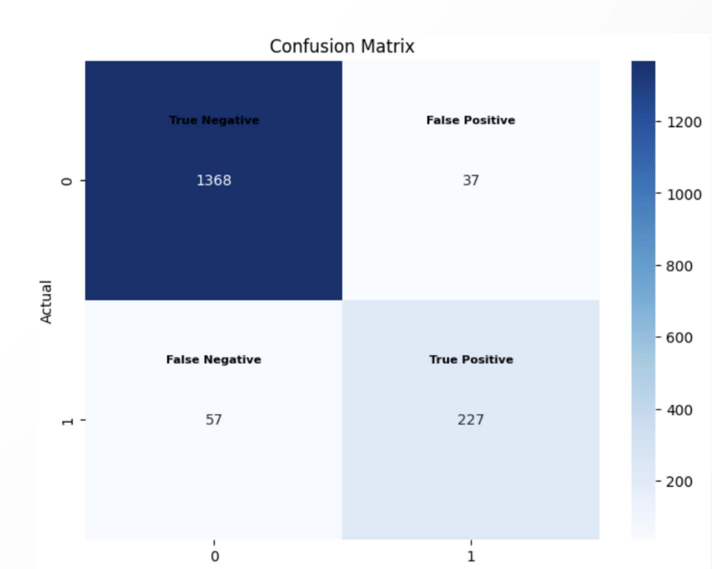
Random Forest ROC AUC: 0.9715

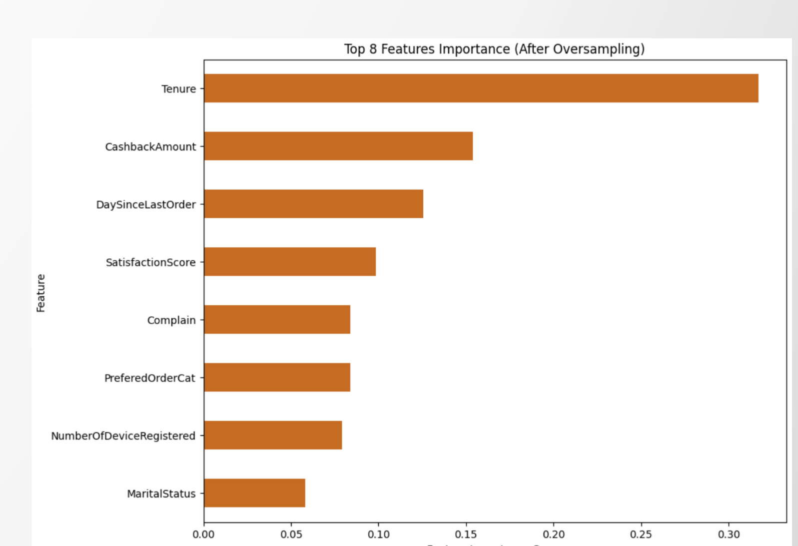
Support Vector Machine ROC AUC: 0.8519

Conclusion after Oversampling









conf_matrix =

[[1368, 37], [57 , 227]]

Tenure the most important feature to predict churn.

Before oversampling:

Random Forest ROC AUC: 0.9715

After oversampling:

Random Forest ROC AUC: 0.9730