E-Commerce Customer Retention Analysis

About the data set

The data set belongs to a leading online E-Commerce company who has a database of customers who have stayed and left. It has 20 columns pertaining to attributes about the customers and 5630 entries referring to each customer.

Business Problem

> An E-Commerce company has a database of their customers who churn and wants to know the attributes that are affecting the customers to leave their platform. These customers will be the focus of their marketing campaign to keep their loyalty and thus increase their sales.

Our Data Set

	count	mean	std	min	25%	50%	75%	max
CustomerID	5630.0	52815.500000	1625.385339	50001.0	51408.25	52815.5	54222.75	55630.0
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.0	0.00	1.0
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.0	16.00	61.0
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.0	3.00	3.0
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.0	20.00	127.0
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.0	3.00	5.0
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.0	4.00	6.0
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.0	4.00	5.0
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.0	6.00	22.0
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.0	1.00	1.0
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.0	18.00	26.0
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.0	2.00	16.0
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.0	3.00	16.0
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.0	7.00	46.0
CashbackAmount	5630.0	177.221492	49.193869	0.0	146.00	163.0	196.00	325.0

Count of Customers who Stayed and Left
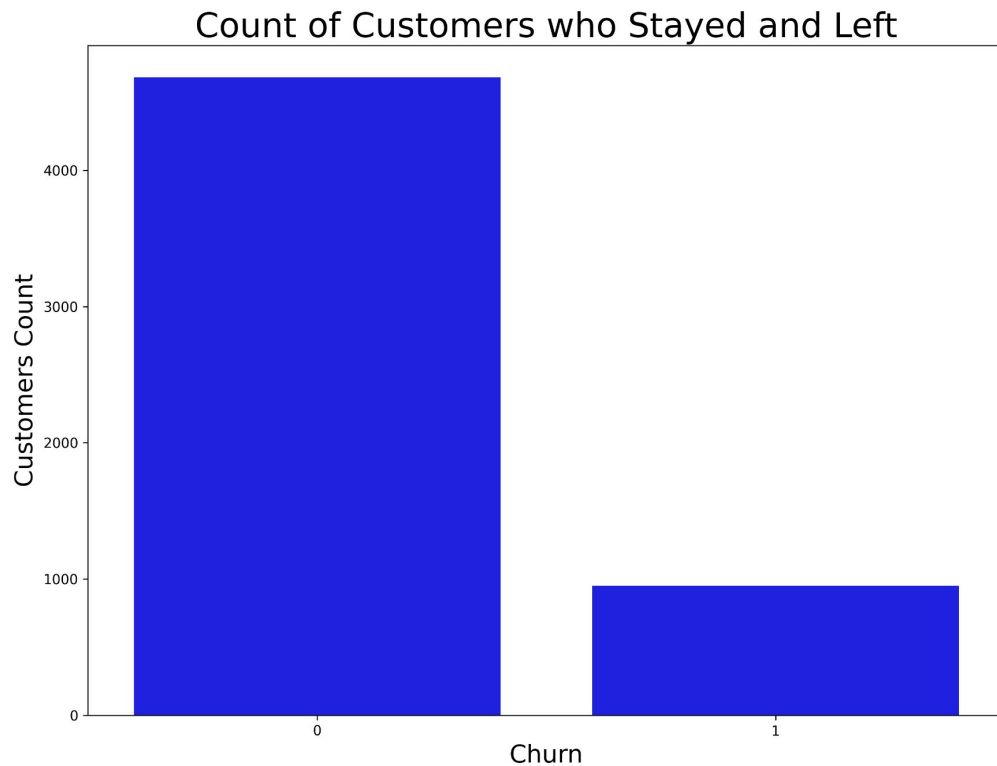
The company currently has a 20% Churn rate

Histogram of Customers Tenure

Count of Customers Based on Their Last Order
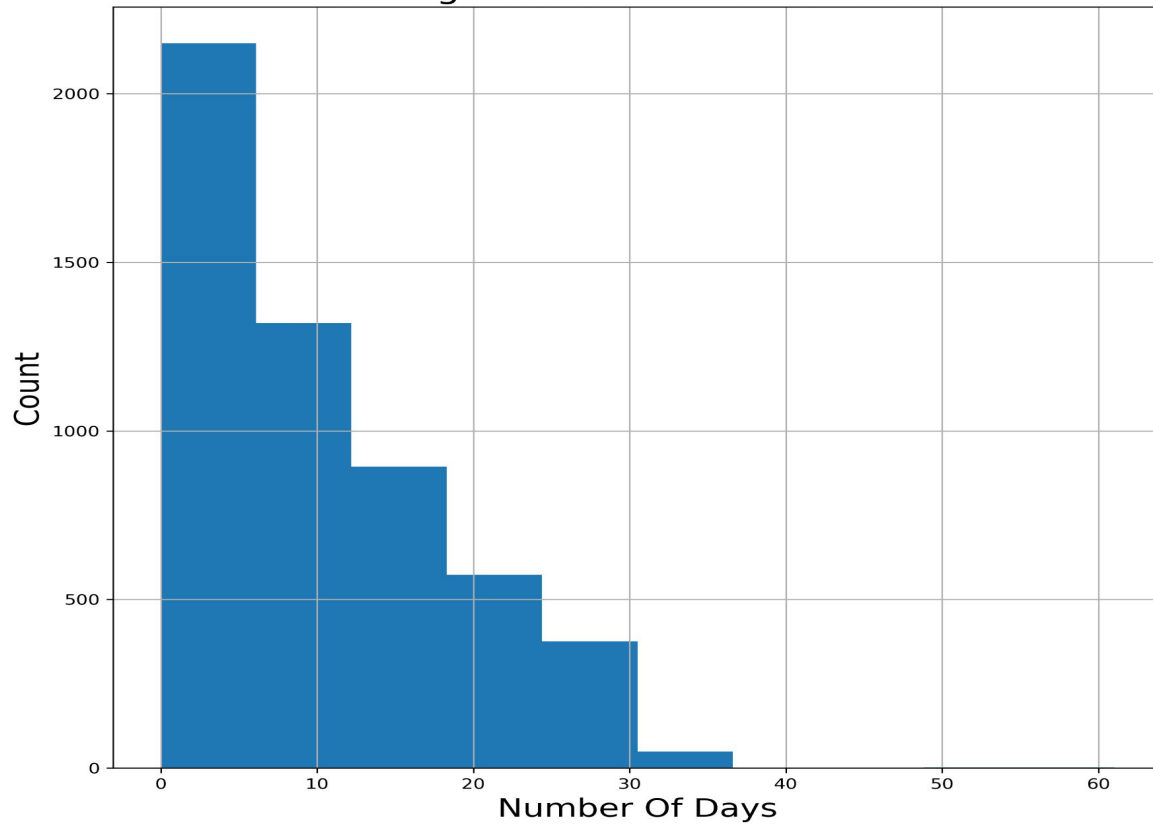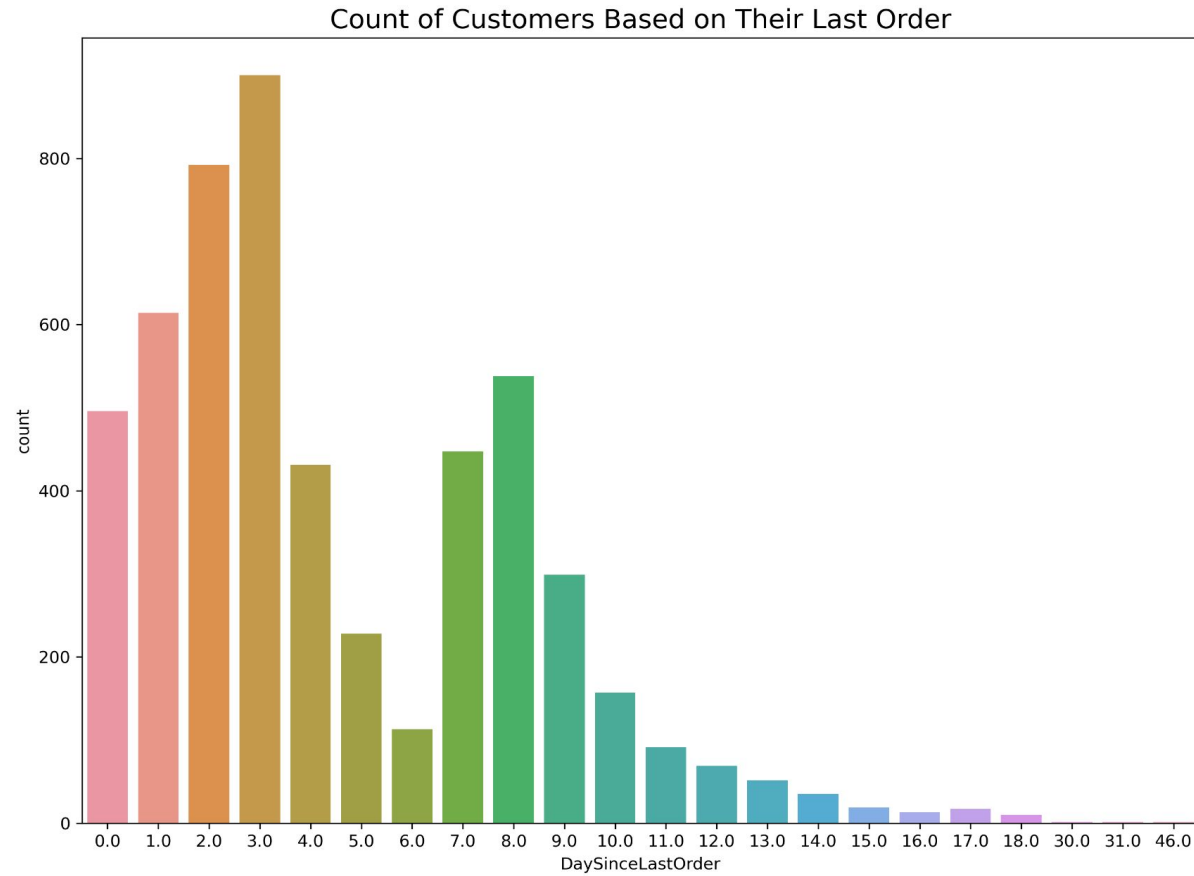
Modelling Technique

Decision Trees Classification

A Decision Tree is a tree (and a type of directed, acyclic graph) in which the nodes represent decisions (a square box), random transitions (a circular box) or terminal nodes, and the edges or branches are binary (yes/no, true/false) representing possible paths from one node to another. The specific type of decision tree used for machine learning contains no random transitions. To use a decision tree for classification or regression, one grabs a row of data or a set of features and starts at the root, and then through each subsequent decision node to the terminal node. The process is very intuitive and easy to interpret, which allows trained decision trees to be used for variable selection or more generally, feature engineering.
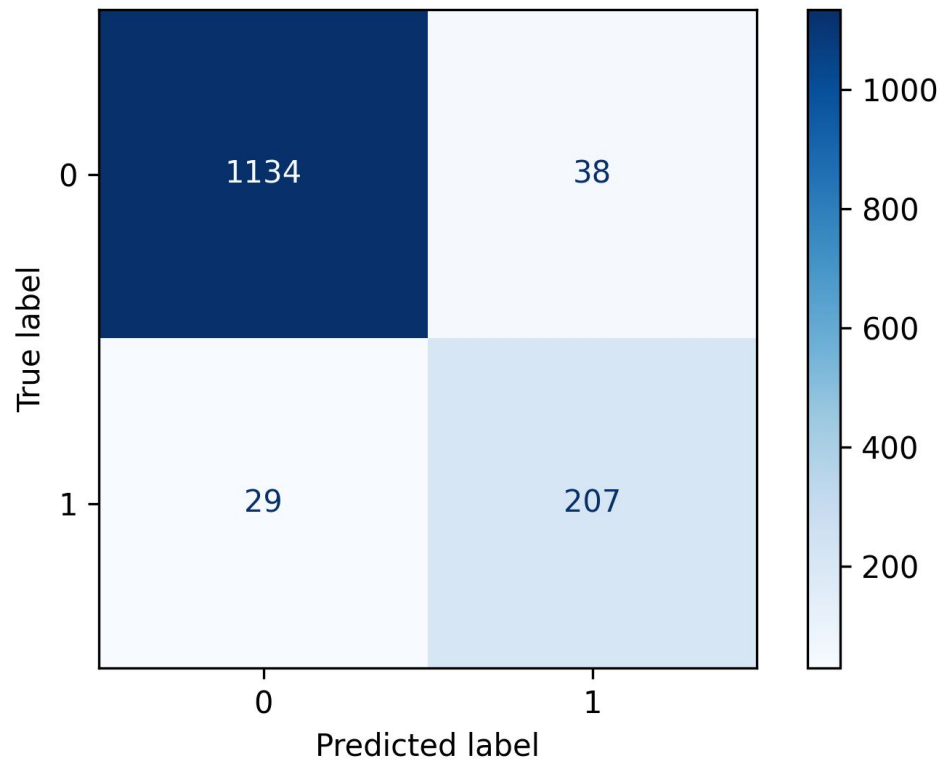
95%

Model Accuracy Score

Confusion Matrix

Index	Feature	Importances
0	Tenure	0.261
17	CashbackAmount	0.085
13	OrderAmountHikeFromlastYear	0.082
3	WarehouseToHome	0.071
16	DaySinceLastOrder	0.07
12	Complain	0.066
11	NumberOfAddress	0.066
9	SatisfactionScore	0.052
4	PreferredPaymentMode	0.042
10	MaritalStatus	0.04
2	CityTier	0.034
8	PreferedOrderCat	0.028
7	NumberOfDeviceRegistered	0.025
1	PreferredLoginDevice	0.024
15	OrderCount	0.02
6	HourSpendOnApp	0.019
5	Gender	0.011
14	CouponUsed	0.005

The features are ordered based on the number of samples affected by splits on those features.
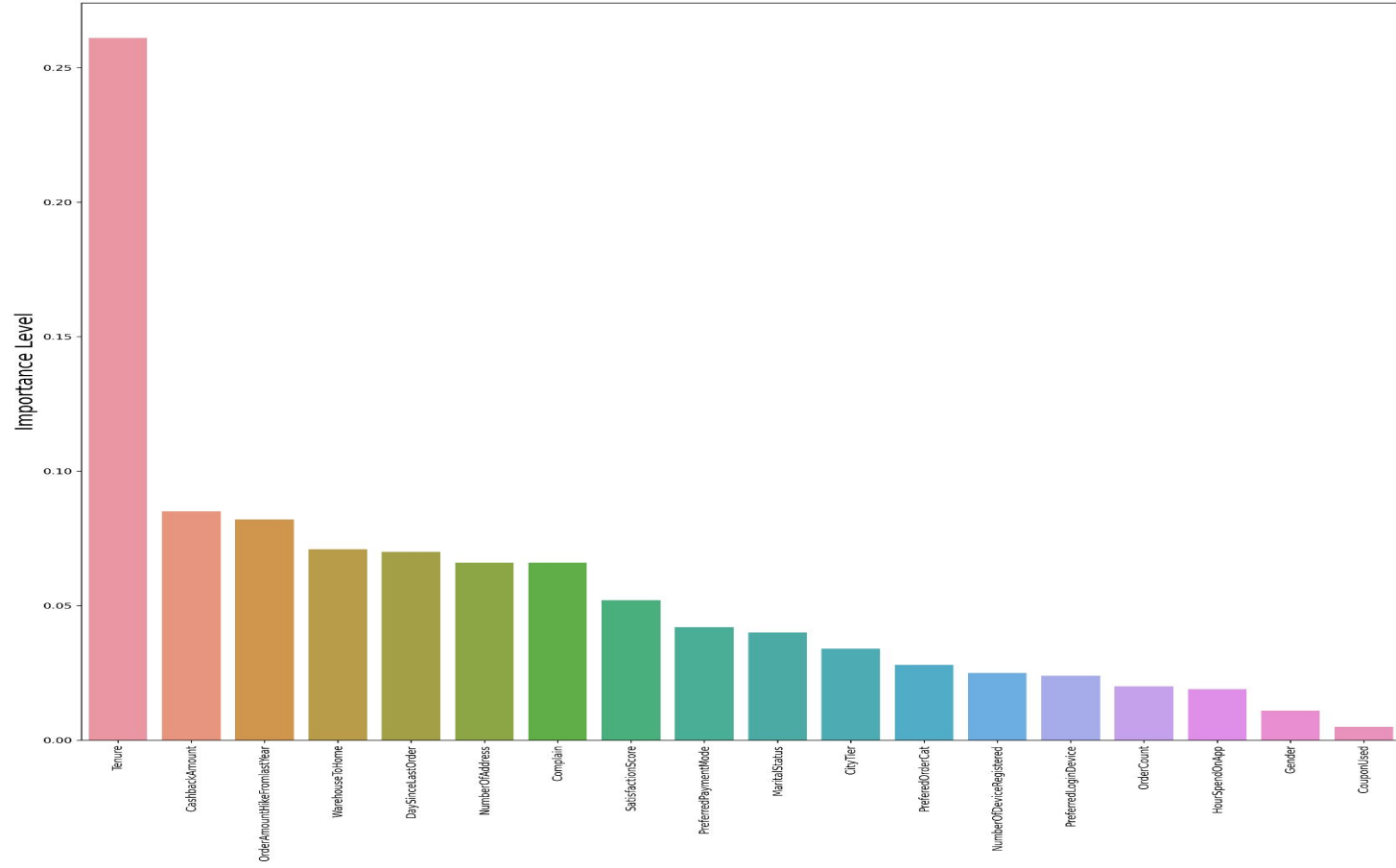
Conclusion

We thoroughly interrogated the data at the outset to gain insight into which features has the highest impact against the Customers who are likely to leave the platform. By utilizing these features, the company can focus marketing on customers who are likely to leave.

More in depth analysis can also be done to further investigate our data.