What?

Utilize customer-level attributes such as Tenure, Cashback Amount, and Warehouse to Home to predict Customer churn on an eCommerce Platform.
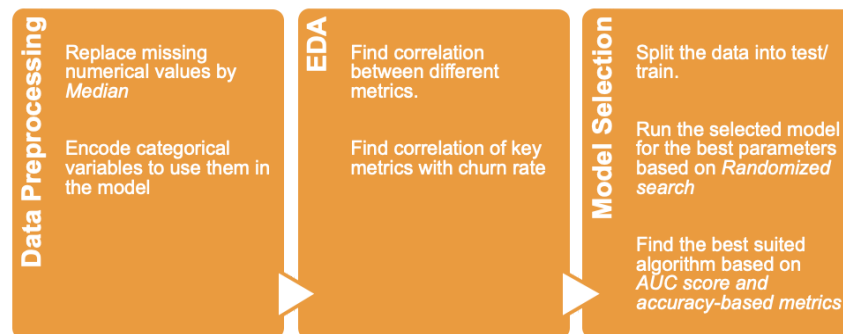
Why?

Customer Churn is the percentage of customers that drop out of the platform during a certain period of time. Predicting if a customer would exit their engagement on the platform helps in creating correct retention strategies to retarget the customers and build consumer intelligence.
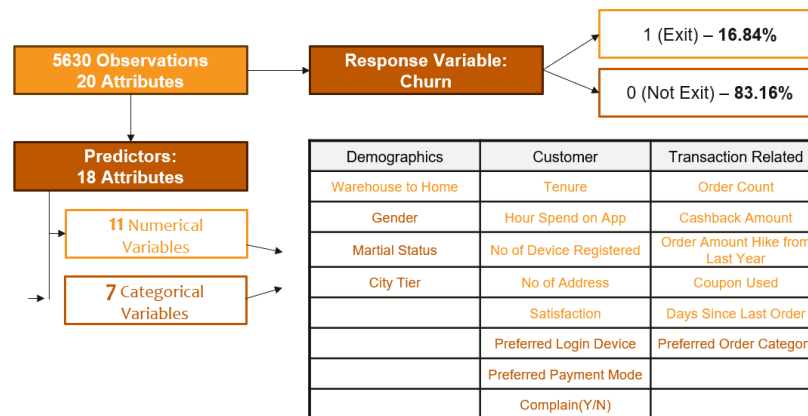
How?

Using the eCommerce dataset, we ran the following Classification models to predict the churn :

1. Logistic Regression
2. Decision Tree
3. KNN
4. Random Forest Search
5. Boosting

Data Preprocessing

Replace missing numerical values by *Median*

Encode categorical variables to use them in the model

EDA

Find correlation between different metrics.

Find correlation of key metrics with churn rate

Model Selection

Split the data into test/ train.

Run the selected model for the best parameters based on *Randomized search*

Find the best suited algorithm based on *AUC score and accuracy-based metrics*

Our dataset consisted of the following variables, out of which five were categorical variables which were converted to numerical after encoding.

5630 Observations
20 Attributes

Response Variable:
Churn

1 (Exit) – **16.84%**

0 (Not Exit) – **83.16%**

Predictors:
18 Attributes

11 Numerical Variables

7 Categorical Variables

| Demographics	Customer	Transaction Related
Warehouse to Home	Tenure	Order Count
Gender	Hour Spend on App	Cashback Amount
Martial Status	No of Device Registered	Order Amount Hike from Last Year
City Tier	No of Address	Coupon Used
	Satisfaction	Days Since Last Order
	Preferred Login Device	Preferred Order Category
	Preferred Payment Mode	
	Complain(Y/N)	

Comparison of the models:

After fitting the models, we compared the AUC score, Precision and Recall for different models to select the optimum model.

Model	ISS Accuracy	OOS Accuracy	Precision	Recall	Parameters
Logistic Regression	0.81	0.83	0.53	0.76	Optimal Threshold = 0.217
Decision Tree Classifier	0.97	0.94	0.91	0.90	Max depth: 10 Min Samples Split: 2 Min Samples Leaf: 1 Criterion: Gini Index
Random Forest Classifier	0.98	0.95	0.99	0.84	N_estimators: 1757 Max depth: 50 Min samples split: 2 Min Samples Leaf: 1 Criterion: Entropy
Gradient Boosting	0.93	0.91	0.81	0.68	N_estimators: 3000 Max depth: 10 Learning Rate: 0.01
KNN Classifier	0.96	0.93	0.90	0.68	K = 4

The key metric that we used to compare the models was Recall. The recall is more important because it's preferable to have a model that does not miss any churns but sometimes classifies "non-churns" as "churns", than a model that does not classify "non-churns" as "churns" but misses a lot of churns. Based on Recall, we are selecting the *Decision tree* as the most optimal model for this problem. The following ROC curve suggests a Decision tree as the best model for our data as well.
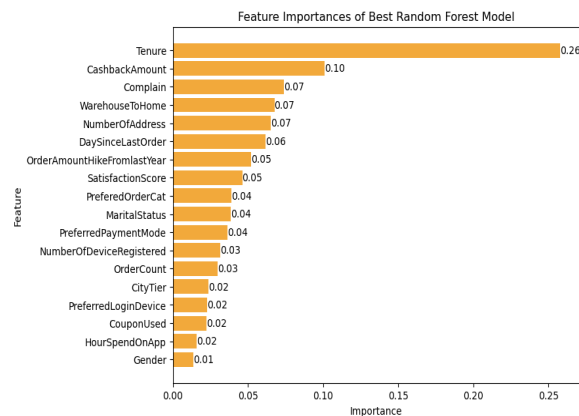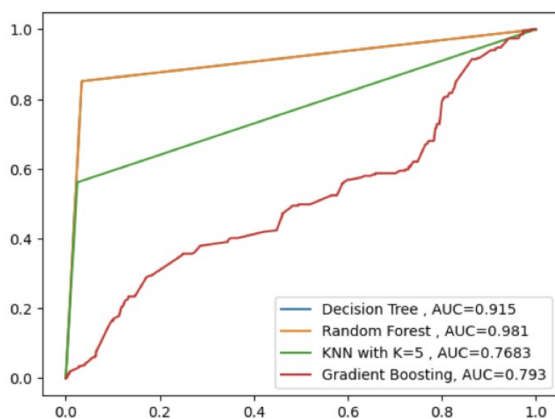
Our data shows that Tenure, Complain, Warehouse to Home distance and cash-back amount are the key variables that dictate a customer's churn from the platform. Although we have chosen the Decision Tree model, similar variable importance is observed in all the other models as well.

Conclusion:

Customers with longer tenure, and farther warehouse-to-home distance are more likely to churn from the platform. Leveraging this information, the platform can devise targeted strategies to retain customers by focusing on customer experience and lower complaint rates.