E-COMMERCE CUSTOMER CHURN

Analysis and Prediction

OUR TEAM

KENSHI PONEVA

Data
Scientist

LUTHFI G. BARKA

Data
Scientist

TOPICS OUTLINE

#1 Problem Formulation

#2 Data Understanding

#3 Findings and Solutions

#4 Conclusions & Recommendations

#5 Questions & Discussions

01

Problem Formulation

Context

As data scientists, we are responsible for a problem of **16,8% customer churn** by identifying the problems, generating insights from data, building ML models with accurate prediction, and providing solutions or recommendations based on the result of analysis and predictions.

Label definition

0 = Customer not churn (loyal customer)

Customer who is not churn is entitled by high numbers of Day Since Last Order, Order Count, and Tenure.

1 = Customer churn (stop/using different app)

Custom who will churn is entitled by low numbers of Day Since Last Order, Order Counts, and Product Category (ie. Gadgets)

PROBLEM UNDERSTANDING

Customer loss

Leads to business growth rate

60-70%

Success rate sales on existing customers

5x Cheaper

Budget on retention cost than acquisition cost

Profit loss

Unhappy Customers = Lost Revenue

5-20%

Success rate sales on new customers

25-90%

Profit gain by a 5% retention rate increase

Source: https://www.profitwell.com/recur/all/customer-acquisition-vs-retention

PROBLEM STATEMENT

GOALS

To predict customer churn as accurate as possible

VALUES

To minimise retention cost
To gain customer lifetime value

Metrics Analysis

Predicted

Actual

	CHURN	NOT CHURN
CHURN	**True Positive (TP)**	**False Negative (FN)**
	Model predicts churn. Actual churn.	Model predicts not churn. Actual churn.
NOT CHURN	**False Positive (FP)**	**True Negative (TN)**
	Model predicts churn. Actual not churn.	Model predicts not churn. Actual not churn.

Type 1 Error (FP)

Consequences: **waste of retention cost** for customer who's loyal already.

Type 2 Error (FN)

Consequences: **losing potential customer** that leads to **CLV & profit loss**.

After understanding the consequences of FP and FN, metrics that we'll use in this project is **f1-score.** Although, we need to seek a **balance between Precision and Recall**, we also need to pay attention on **Recall** score**.** Besides, we use f1-score because there is an **uneven class distribution** (large number of Actual Negatives).

DATA UNDERSTANDING 02

ABOUT DATA

	Column Name	Data Type	Data Count	Missing Value	Missing Value %	Number of Unique	Unique Sample
0	Churn	int64	5630	0	0.00	2	[1, 1]
1	CityTier	int64	5630	0	0.00	3	[2, 2]
2	Complain	int64	5630	0	0.00	2	[1, 1]
3	NumberOfAddress	int64	5630	0	0.00	15	[8, 6]
4	SatisfactionScore	int64	5630	0	0.00	5	[4, 2]
5	NumberOfDeviceRegistered	int64	5630	0	0.00	6	[2, 1]
6	OrderCount	float64	5630	258	4.58	16	[12.0, 6.0]
7	CouponUsed	float64	5630	256	4.55	17	[1.0, 2.0]
8	OrderAmountHikeFromlastYear	float64	5630	265	4.71	16	[11.0, 20.0]
9	CashbackAmount	float64	5630	0	0.00	2586	[146.41, 127.19999999999999]
10	HourSpendOnApp	float64	5630	255	4.53	6	[2.0, 1.0]
11	WarehouseToHome	float64	5630	251	4.46	34	[18.0, 28.0]
12	Tenure	float64	5630	264	4.69	36	[1.0, 23.0]
13	DaySinceLastOrder	float64	5630	307	5.45	22	[13.0, 2.0]
14	MaritalStatus	object	5630	0	0.00	3	[Single, Married]
15	Gender	object	5630	0	0.00	2	[Female, Male]
16	PreferredPaymentMode	object	5630	0	0.00	7	[Cash on Delivery, Cash on Delivery]
17	PreferredLoginDevice	object	5630	0	0.00	3	[Phone, Computer]
18	PreferedOrderCat	object	5630	0	0.00	6	[Fashion, Fashion]

ABOUT DATA

SOURCE & CONTRIBUTOR

DBA - Ankit Verma

Published date

January 2021

Shape

5630 observations
20 features

Data type

5 categorical
15 numerical

Features Quality

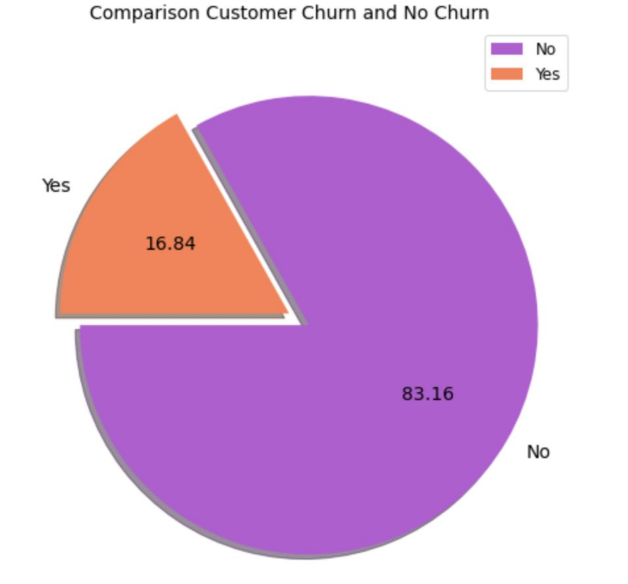
Not normally distributed, outlier in few numerical features, missing values.

FINDINGS AND SOLUTIONS

03

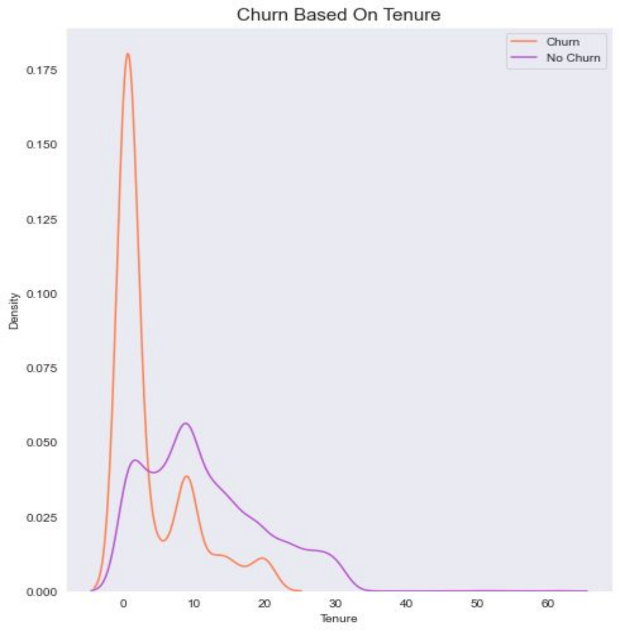
DATA ANALYSIS

COMPARISON OF CUSTOMER CHURN AND NO CHURN



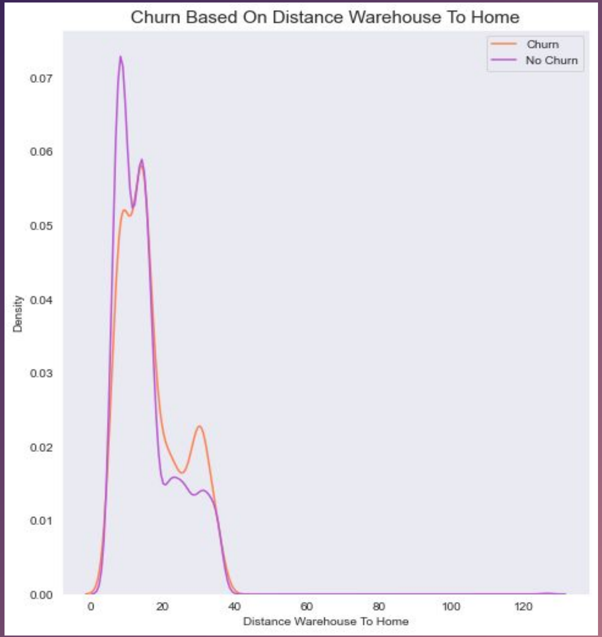
16.84% or 948 customers have churned from total customers

Churn Based On Tenure



Customers with tenure less than 3 months tend to churn.

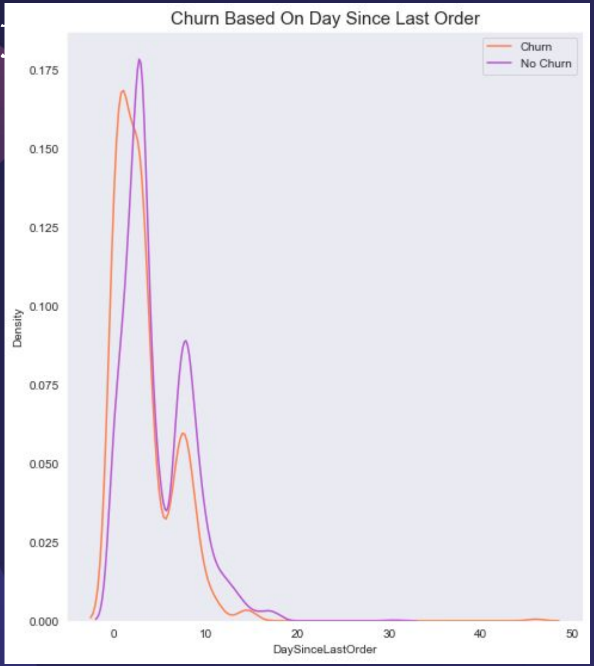
Churn Based On Distance Warehouse To Home



The further away warehouse to home, more customers tend to churn

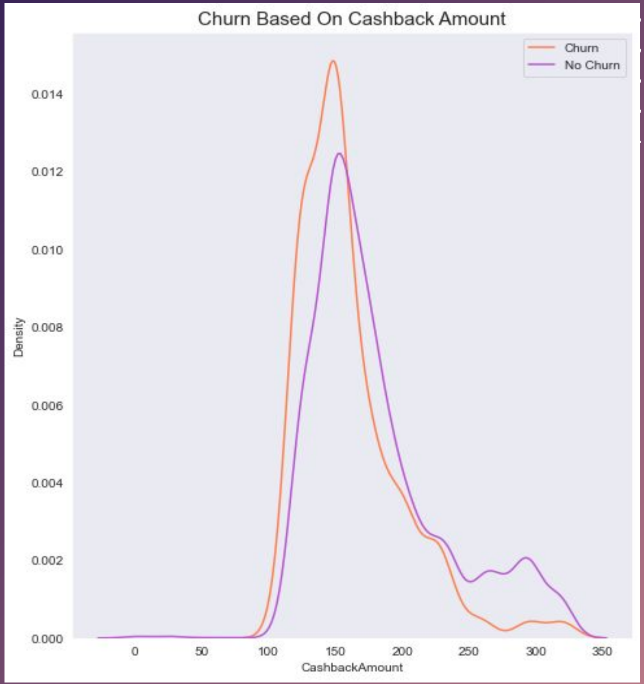
A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analagous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions.

Churn Based On Day Since Last Order

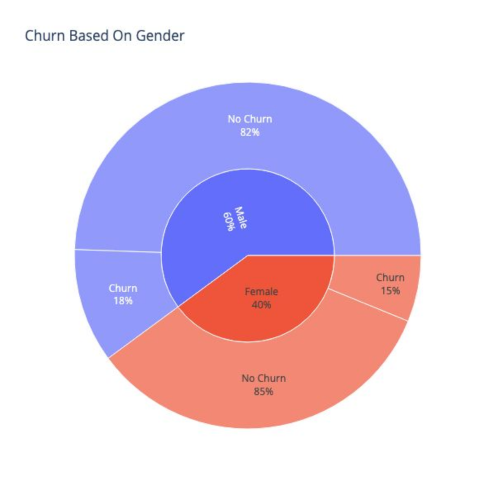


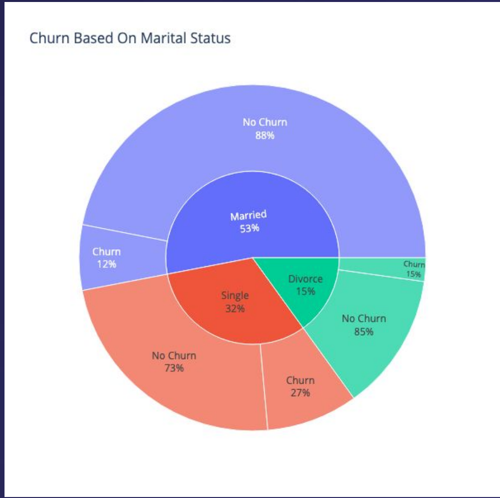
Mostly customers who make new orders decide to churn

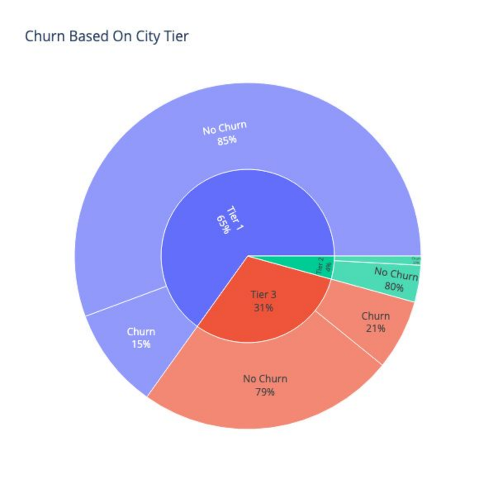
Churn Based On Cashback Amount



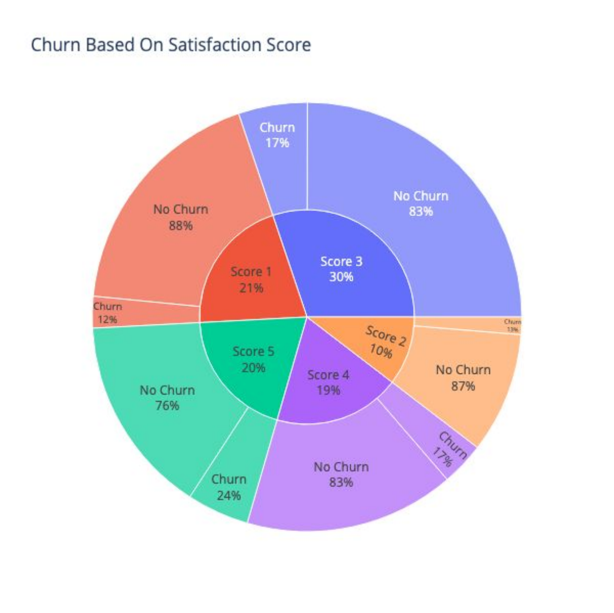
The less cashback amount customers get, the higher risk of customers churn

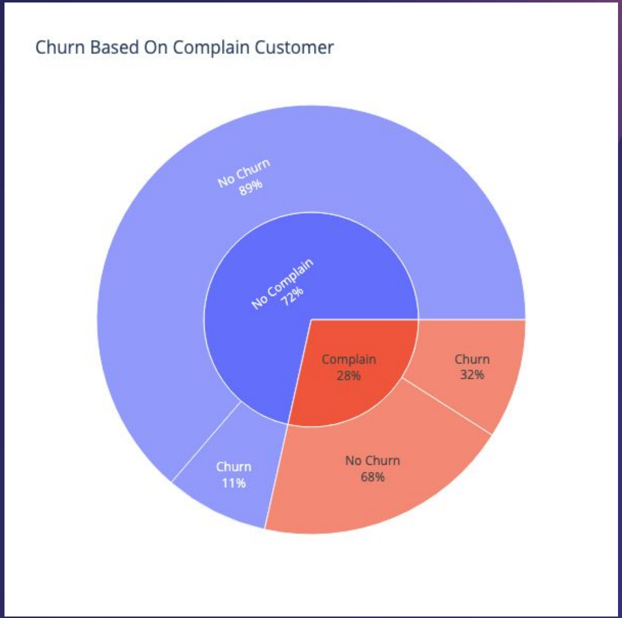






- ➔ Male customers churn more than female
- ➔ Single customers have the highest churn rate on marital status
- ➔ The higher city tier, the more customers tend to churn





- ➔ Mostly customers give 3 score on app. The higher score, the higher customers churn rate.
- ➔ More customers doing complain, more customers churn

MACHINE LEARNING MODEL

Machine Learning Steps

#1 Data Splitting

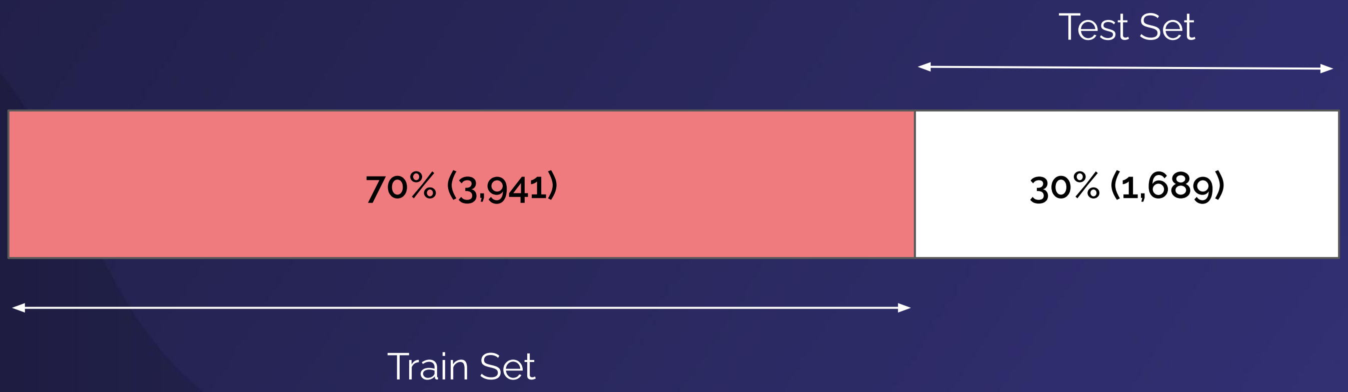
#2 Feature Engineering

#3 Model Evaluation & Selection

#4 Hyperparameter Tuning

#5 Feature Importance

DATA SPLITTING



FEATURE ENGINEERING

MISSING VALUES

- 7 features (float type).
- Using Iterative Imputer technique.

Anticipate categorical features from unseen data using strategy 'constant'.

NORMALISATION

Transforming numerical data into the range [0,1].

- 6 numerical features using robust scaler.
- One numerical feature using minmax scale.

ENCODING

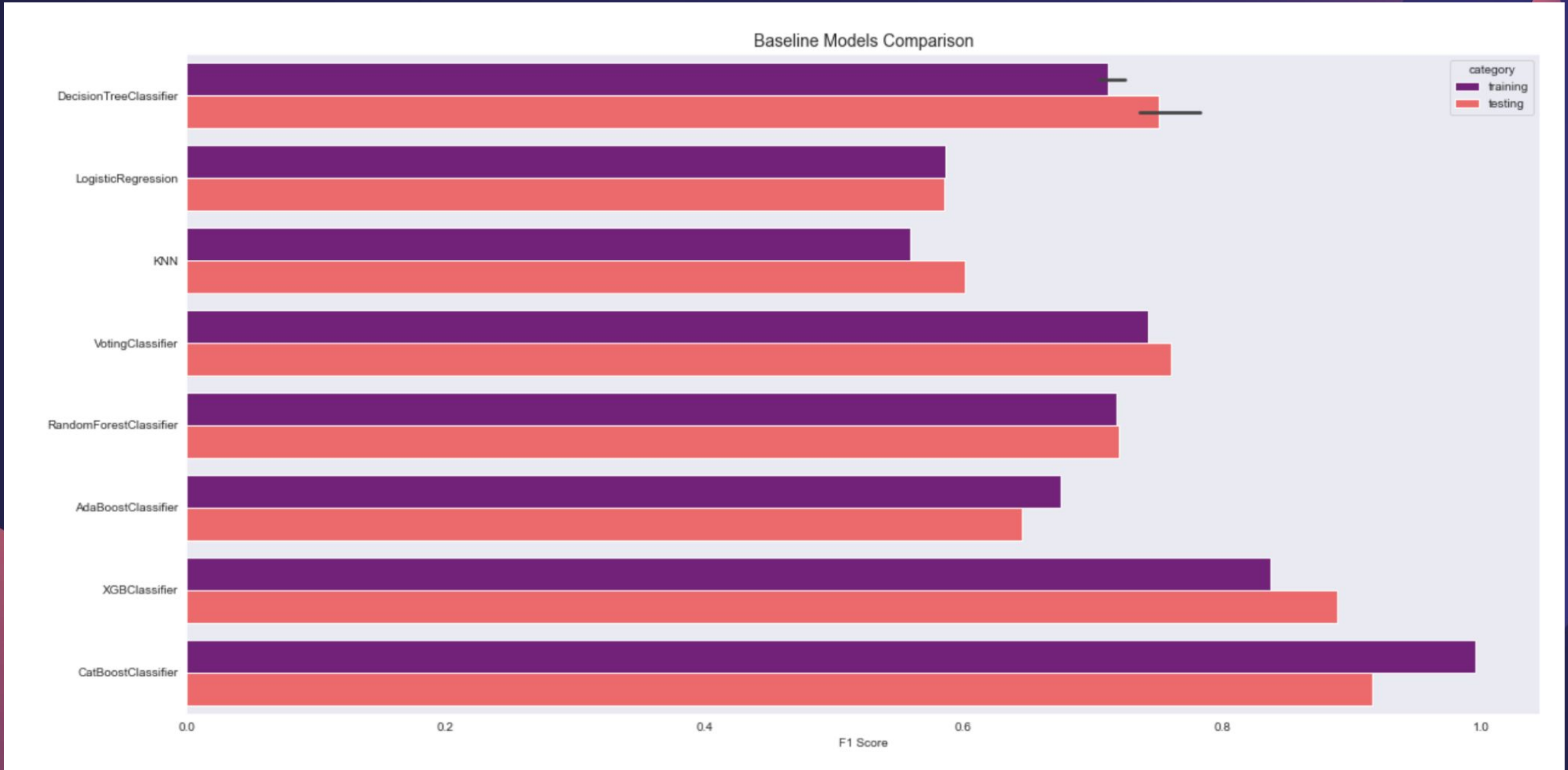
Transforming categorical data into numerical.

- Using OneHot Encoding for all categorical features.

MODEL SELECTION

	PRECISION	RECALL	F1	WEAK POINT
DECISION TREE	61%	92%	74%	Low precision
LOGISTIC REGRESSION	44%	86%	59%	Low precision
KNN	74%	50%	60%	Low recall
VOTING	67%	87%	76%	Low precision
RANDOM FOREST	64%	83%	72%	Low Precision
ADABOOST	77%	56%	64%	Low recall
XGBOOST	93%	85%	89%	None
CATBOOST	95%	88%	91%	None

MODEL VISUALISATION



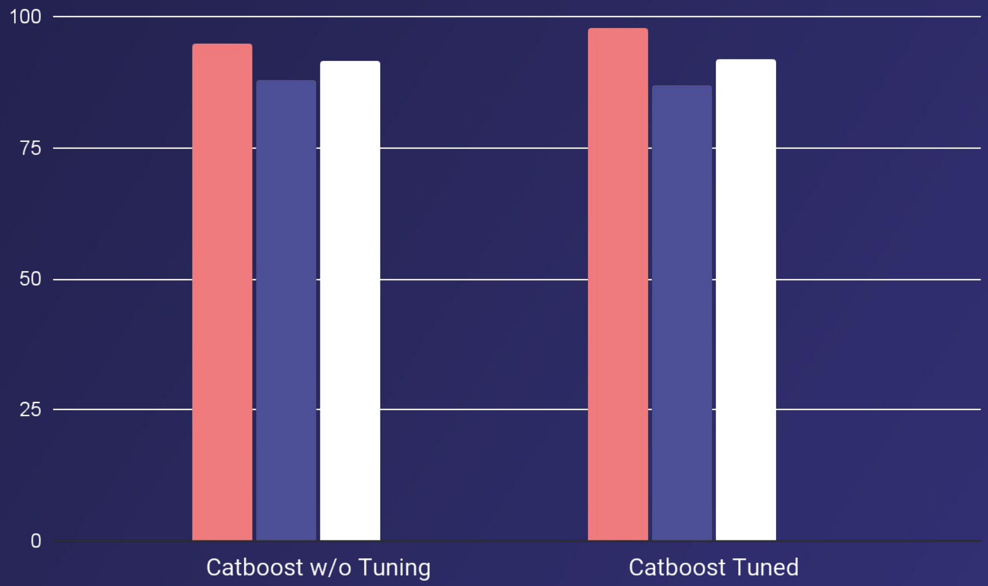
CATBOOST BEFORE & AFTER HYPERPARAMETER TUNING

PRECISION

Before: 95%
After: 98%

RECALL

Before: 88%
After: 87%



F1-SCORE

Before: 91.5%
After: 92%

PARAMETERS TUNING

Before Tuning

- Iterations: 1000
- Eval metric: F1
- Random state

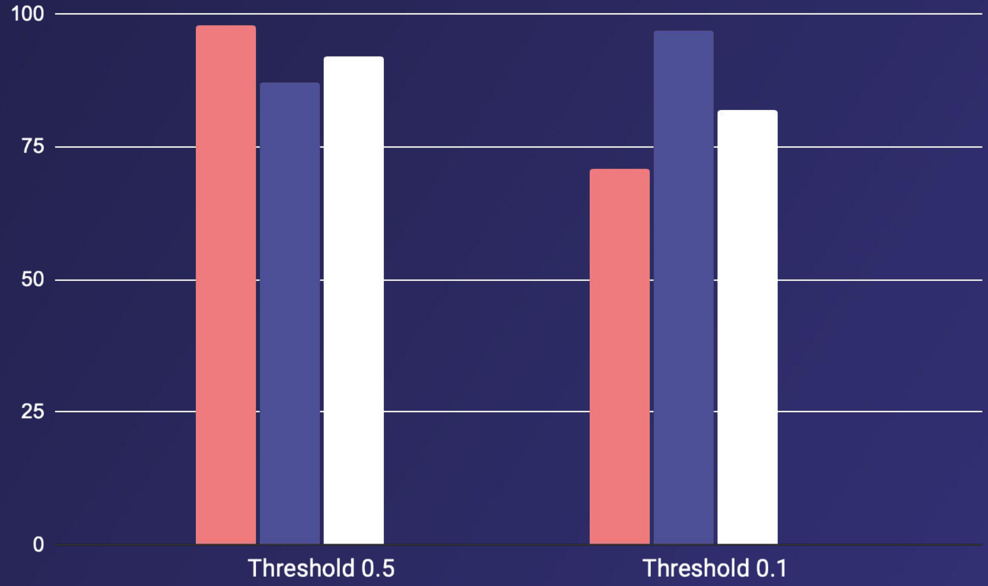
After Tuning

- Iterations: 1500,
- Eval metric: F1,
- Depth: 10,
- L2 leaf reg: 5,
- Learning rate: 0.01,
- Od type: Iter,
- Od wait: 100,
- Random state

CATBOOST TUNED - THRESHOLD ADJUSTED

PRECISION
Before: 98%
After: 71%

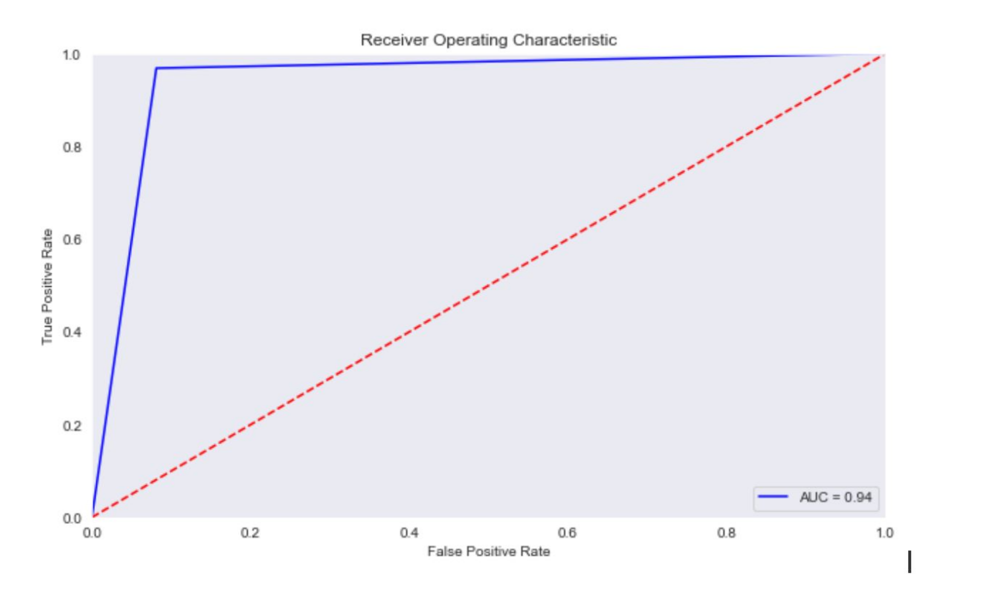
RECALL
Before: 87%
After: 97%



F1-SCORE
Before: 91.5%
After: 82%

We want to increase Recall score to lower False Negative rate by changing the threshold to be 0.1.

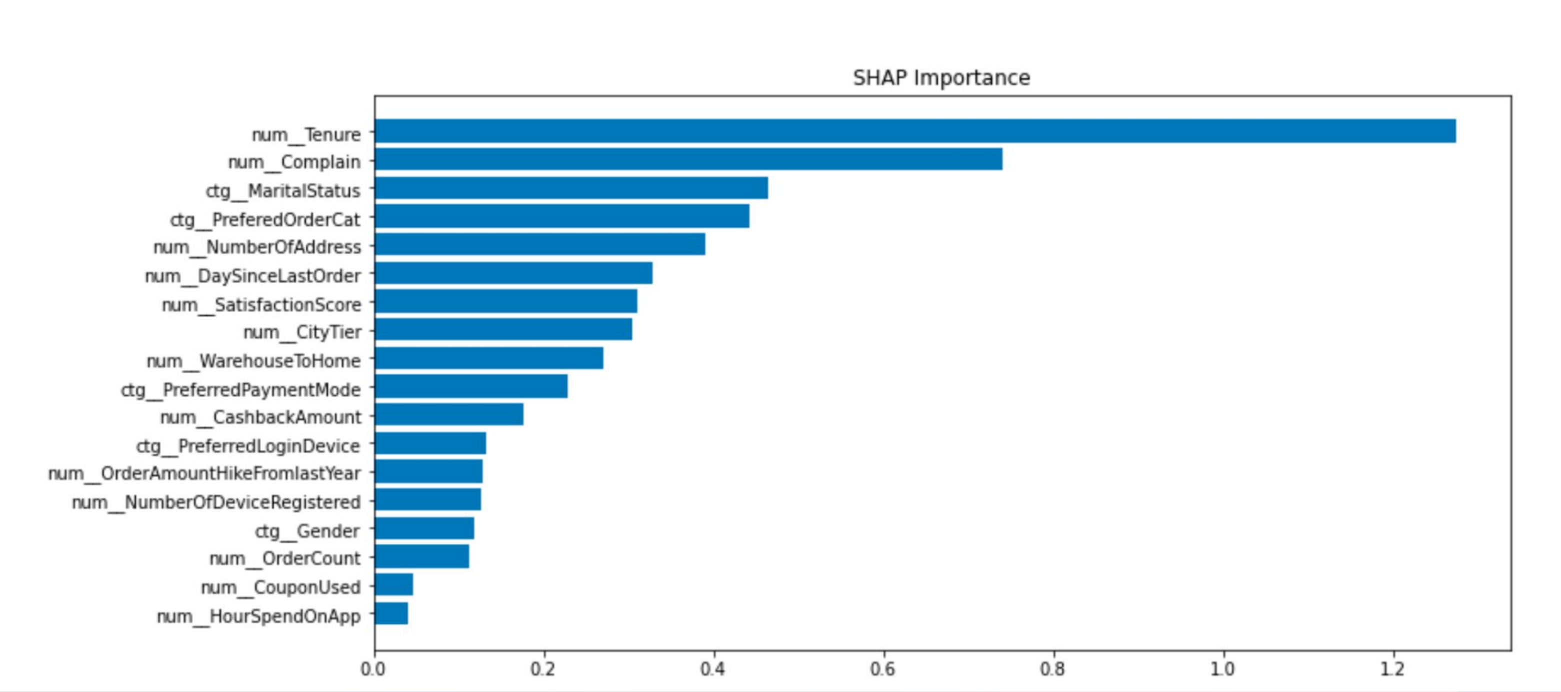
ROC-AUC SCORE



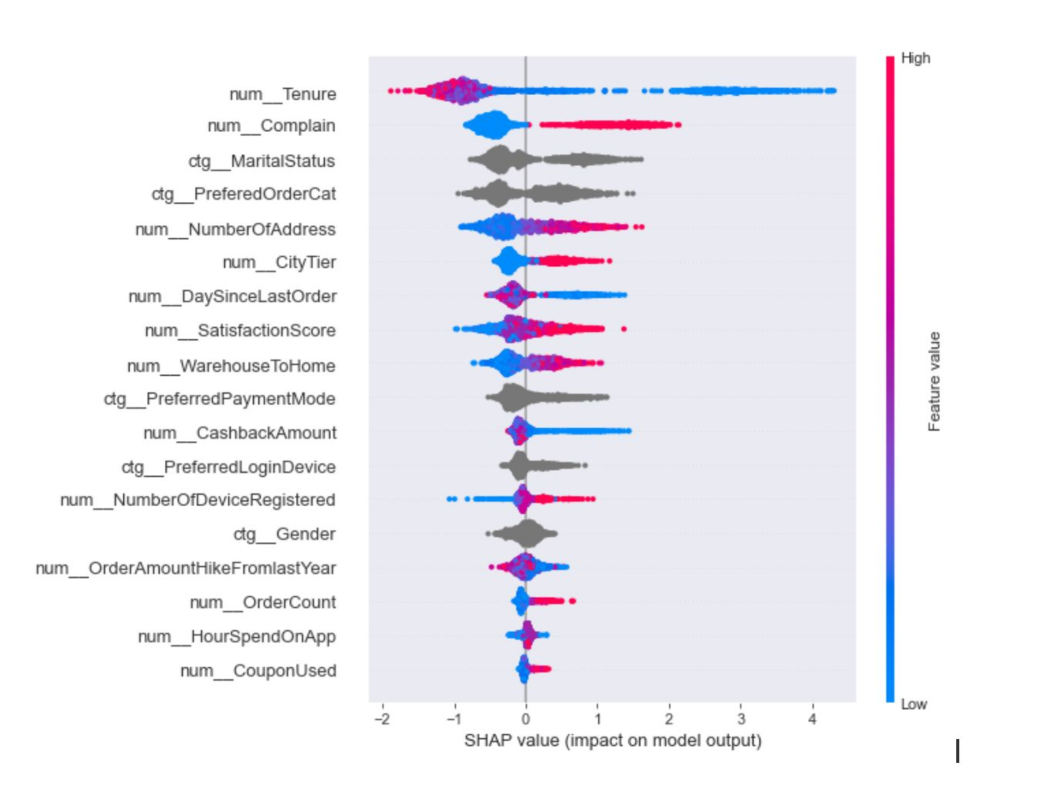
- Probability curve to plot TPR and FPR.
- AUC is a measure for the model to distinguish between classes.
- Higher the AUC, the better the performance of the model.

Our model has ROC AUC score at 0.94 which means that our model has an outstanding ability to distinguish between customer who will churn or who will retain.

FEATURE IMPORTANCE

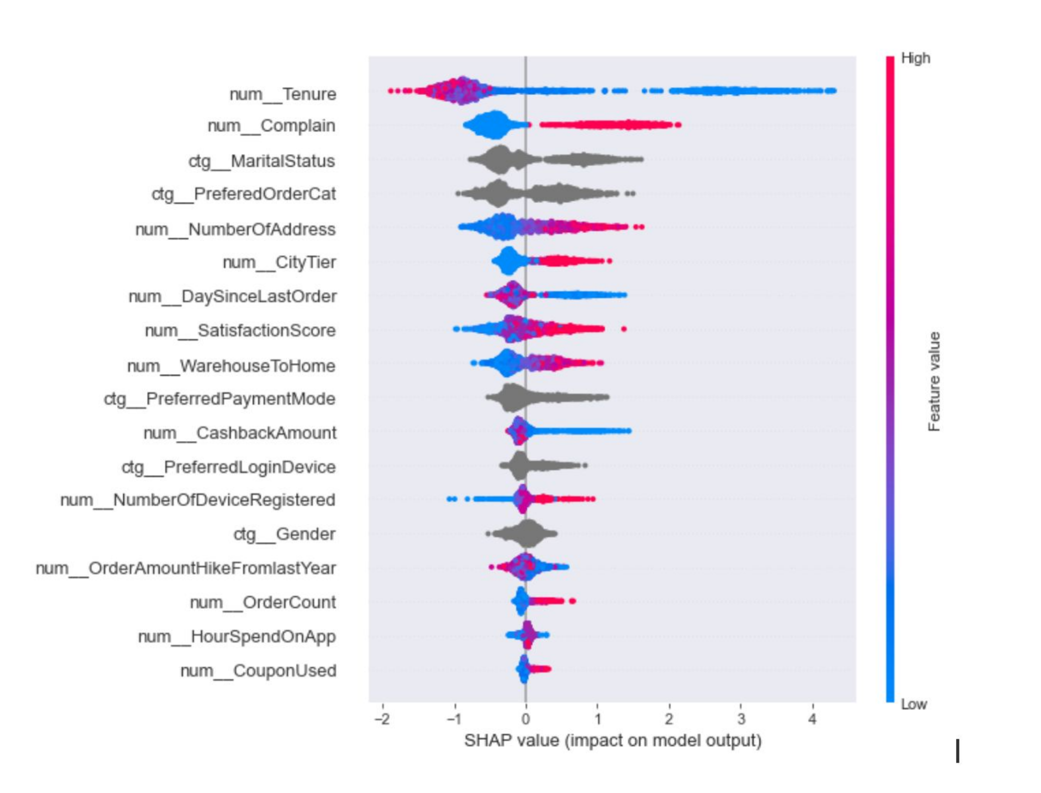


SHAP ANALYSIS



1. **Tenure**
   The shorter customer's tenure, the higher risk of churn.
2. **Complain**
   The more complaints from customer, the higher churn is.
3. **Number of Address**
   The more addresses customers registered, the higher churn is.
4. **Day Since Last Order**
   The smaller interval of order, the higher churn is.

SHAP ANALYSIS



5. **Satisfaction Score**
   The higher rate given, the higher risk of churn.

6. **City Tier**
   The higher tier level, the higher risk of churn.

7. **Warehouse to Home**
   The longer distance between warehouse to customer's home, the higher risk of churn.

8. **Cashback Amount**
   The smaller amount of cashback received by a customer, the higher risk of churn.

04

CONCLUSION AND RECOMMENDATION

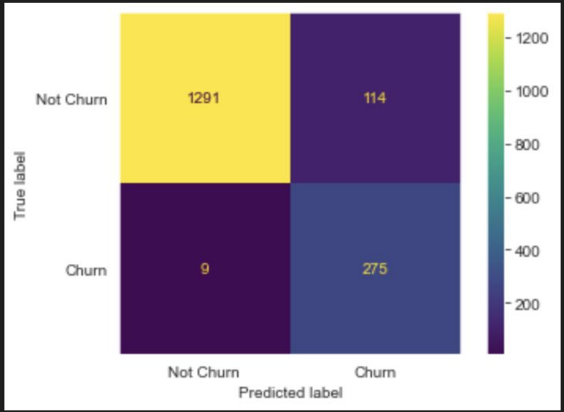
TECHNICAL CONCLUSION

```
CLASSIFICATION REPORT - CATBOOST TUNED:
              precision    recall  f1-score   support

           0       0.99      0.92      0.95      1405
           1       0.71      0.97      0.82       284

    accuracy                           0.93      1689
   macro avg       0.85      0.94      0.89      1689
weighted avg       0.94      0.93      0.93      1689
```

CONFUSION MATRIX - CATBOOST TUNED



Insights

Our model scored 82% of f1-score

- There is 26% customer not churn predicted will churn (Recall 1 - Precision 1).

Conclusion by Recall:

- Able to detect 97% of customers who will churn (recall 1)
- Finds 92% of customers who'll retain/stay (recall 0).

Conclusion by Precision:

- Predictive accuracy of customers who do not churn reached 71% (precision 1).

BUSINESS CONCLUSION

Let's say, the retention cost (ie. 50% discount offer) per customer is $20 and we currently generated 200 customer which churn 100 people and not churn 100 people. The calculation as follows:

Without Model:

- **Total Cost:** 200 x $20 = USD 4,000
- **Customer churn w/ disc:** 100 people
- **Customer churn w/o disc:** 0 people
- **Budget waste =>** 100 x 20 USD = USD 2000 (it's a waste of cost because 100 who retain will accept again this offer that's already loyal)
- **Cost saving =>** USD 0

With Model:

- **Total Cost:** (97 x 20 USD) + (26 x 20 USD) = 1940 USD + 520 USD = USD 2460
- **Customer churn w/ disc:** 97 people
- **Customer churn w/o disc:** 3 people
- **Budget waste** => 26 x $20 = USD 520
- **Cost saving**=> 92 x $20 = USD 1840

Therefore, by using this model we can help Marketing team to align with our goals **to minimise retention cost** because our model is able to minimise budget waste; and **to increase Customer Life Value** because our model can predict customer who will churn about 97 %.

RECOMMENDATION - MODEL

There are few things we can optimise the model performance by following recommendations below:

- **Improve the database system** on the application or website so that customer activities can be saved automatically in aiming to **reduce missing values** for ‘real time’ features such as \`HourSpendOnApp\`, \`DaySinceLastOrder\`, \`OrderAmountHikeFromlastYear\`, \`CouponUsed\`, \`OrderCount\`.

- **Adding new features or columns** that may add to customer information by sending online surveys via email along with a 'reward' like voucher discount once the survey has been submitted.

- **Try another ML algorithms** with optimal performance or **reset hyperparameter tuning** on the primary model. Especially if there are seasonal campaigns such as numbers of discount offers on special days.

- **Analyse the data** in the sense of **wrong prediction** by our model to find out the reason behind it and its characteristics.

RECOMMENDATION - FEATURE IMPORTANCE

Tenure	Incentive rewards to customer with 0-1 year tenure. This small actions can retain customer who'll churn almost up to 15%.
Complain	Optimise service quality in various digital channels to provide customers a better feedback.
City Tier	Scale up business infrastructure in the area of city tier 2 and tier 3 to increase potential leads and gain market share.
Day Since Last Order	Launch a new program such as 30-day free trial subscription on exclusive product/services to create curiosity on customers.
Warehouse to Home	Offer a low rate/free delivery cost to customers who live far away from the local warehouse.

THANK YOU


The easiest way to
launch your online store
Designed with you in mind
MacBook Pro

RESOURCES

Notebook
https://github.com/kponeva/ecommerce-customer-churn-prediction

Articles
https://www.profitwell.com/recur/all/customer-acquisition-vs-retention
https://www.huify.com/blog/acquisition-vs-retention-customer-lifetime-value
https://blog.hubspot.com/service/how-to-reduce-customer-churn
https://blog.hubspot.com/service/customer-retention-strategies