"Know Your Customers"
Customer Churn Analysis & Prediction

Ishwarya Nikitha Nischita
San Diego State University

CS577, Fall 2023

Problem Statement

Develop a predictive model using ML algorithms for analyzing customer attrition in the E-commerce industry to identify the customers at risk and to improve customer retention which would thereby increase revenue of the business.

Background

- Existing research utilizes k-means clustering and Support Vector Machines to test the predictive ability of the model.
- The research doesn't include testing the predictive ability of the customer churn rate using the most common prediction algorithm that is Decision Tree.

Project Objective

- Identify the key features that affect E-comm customer attrition.
- Identify the optimal algorithm from Logistic Regression, Decision Tree and Random Forest that can accurately predict customer churn based on evaluation of performance metrics.
- Provide customer retention strategies based on insights gained from the analysis.

Data Source

- Data set is collected from Kaggle.
- Data set has 20 columns and 5630 rows or records.

Proposed Approach

Data Set: Kaggle (E-commerce dataset)

↓

Data pre-processing

↓

Modeling: Logistic Regression, Random Forest and Decision Tree

↓

Analysis of Results

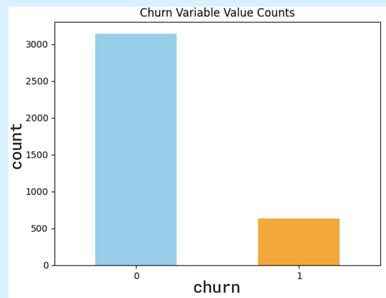
Performance comparison of prediction model between Logistic Regression, Random Forest and Decision Tree

Important Features - Random Forest

Feature	Importance
Tenure	0.243432
WarehouseToHome	0.075840
CashbackAmount	0.070549
NumberOfAddress	0.064670
DaySinceLastOrder	0.058423
OrderAmountHikeFromlastYear	0.047792
Complain	0.045635
NumberOfDeviceRegistered	0.043693
SatisfactionScore	0.042697
OrderCount	0.032452
PreferedOrderCat_Laptop & Accessory	0.028947
CouponUsed	0.024885
HourSpendOnApp	0.021961
CityTier	0.021445
MaritalStatus_Single	0.020153
Gender	0.018563
MaritalStatus_Married	0.017441
PreferredLoginDevice_Computer	0.017343
PreferredPaymentMode_Debit Card	0.014309
PreferedOrderCat_Mobile Phone	0.013111
PreferredPaymentMode_Credit Card	0.012634
PreferredLoginDevice_Mobile Phone	0.011004
PreferredLoginDevice_Phone	0.010825
PreferredPaymentMode_E wallet	0.009845
PreferredPaymentMode_COD	0.008503
PreferedOrderCat_Fashion	0.008144
PreferredPaymentMode_UPI	0.005524
MaritalStatus_Divorced	0.005130
PreferedOrderCat_Mobile	0.001858
PreferredPaymentMode_CC	0.001580
PreferedOrderCat_Others	0.000828
PreferredPaymentMode_Cash on Delivery	0.000786
PreferedOrderCat_Grocery	0.000000

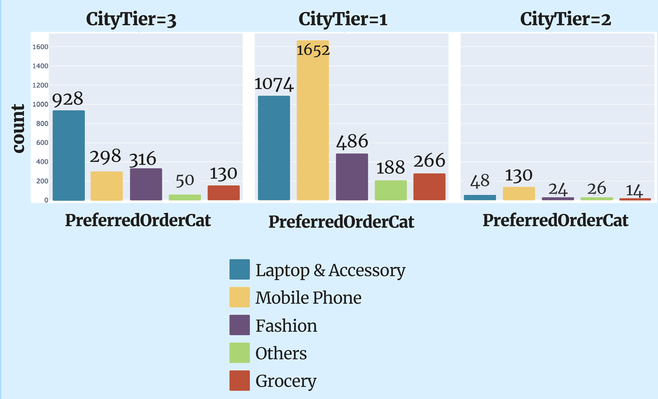
Visualization

Value Counts of 'Churn' Variable

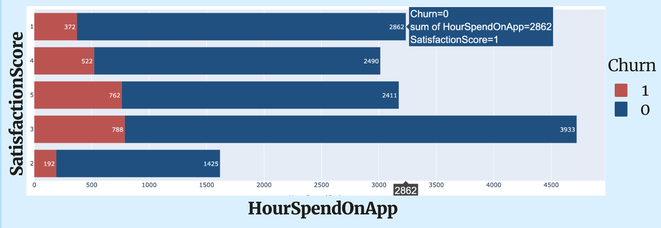


	Count
0	3143
1	631

City Tier vs PreferredOrderCat



City Tier vs PreferredOrderCat

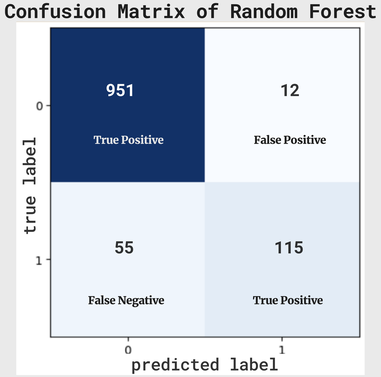


Experimental Results

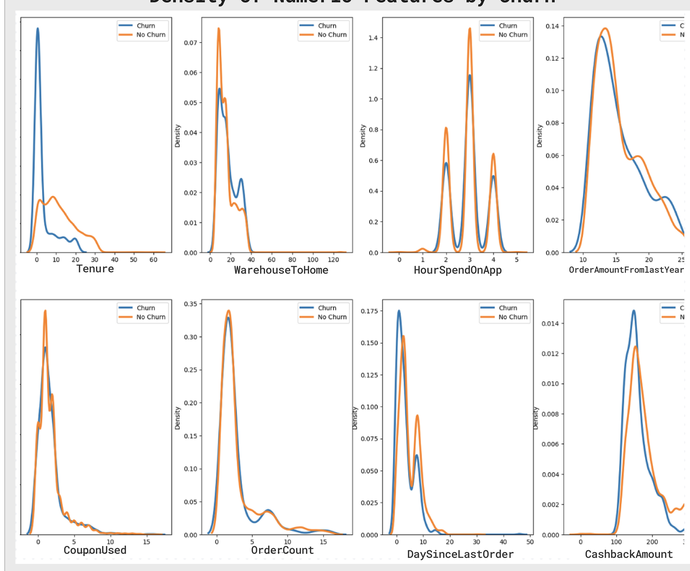
Comparison Table of Models

	Random Forest	Pruned Decision Tree	Logistic Regression
Accuracy	0.94086496	0.910856134	0.903795234
Precision	0.905511811	0.734693878	0.765217391
Recall	0.676470588	0.635294118	0.517647059
F1-Score	0.774410774	0.681388013	0.61754386

Confusion Matrix of Random Forest



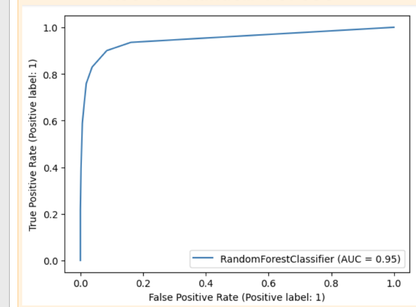
Density of Numeric Features by Churn



Evaluation

- The Random Forest model has a higher accuracy (94.09%) compared to the pruned Decision Tree (91.09%).
- The Random Forest model also has a higher precision among all the other models.

ROC of Random Forest



AUC = 0.95

Conclusion

- The company should focus on less expensive and less cheaper products for the city tier 2 people.
- Random Forest performs better than Decision Tree in terms of accuracy and robustness, especially when dealing with categorical features, high-dimensional spaces, and large datasets.
- Hyper Parameter Tuning can increase the accuracy of the model

References

- Xiahou, X., & Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research