CAPSTONE PROJECT

Airbnb App User Churn Comprehensive Analysis

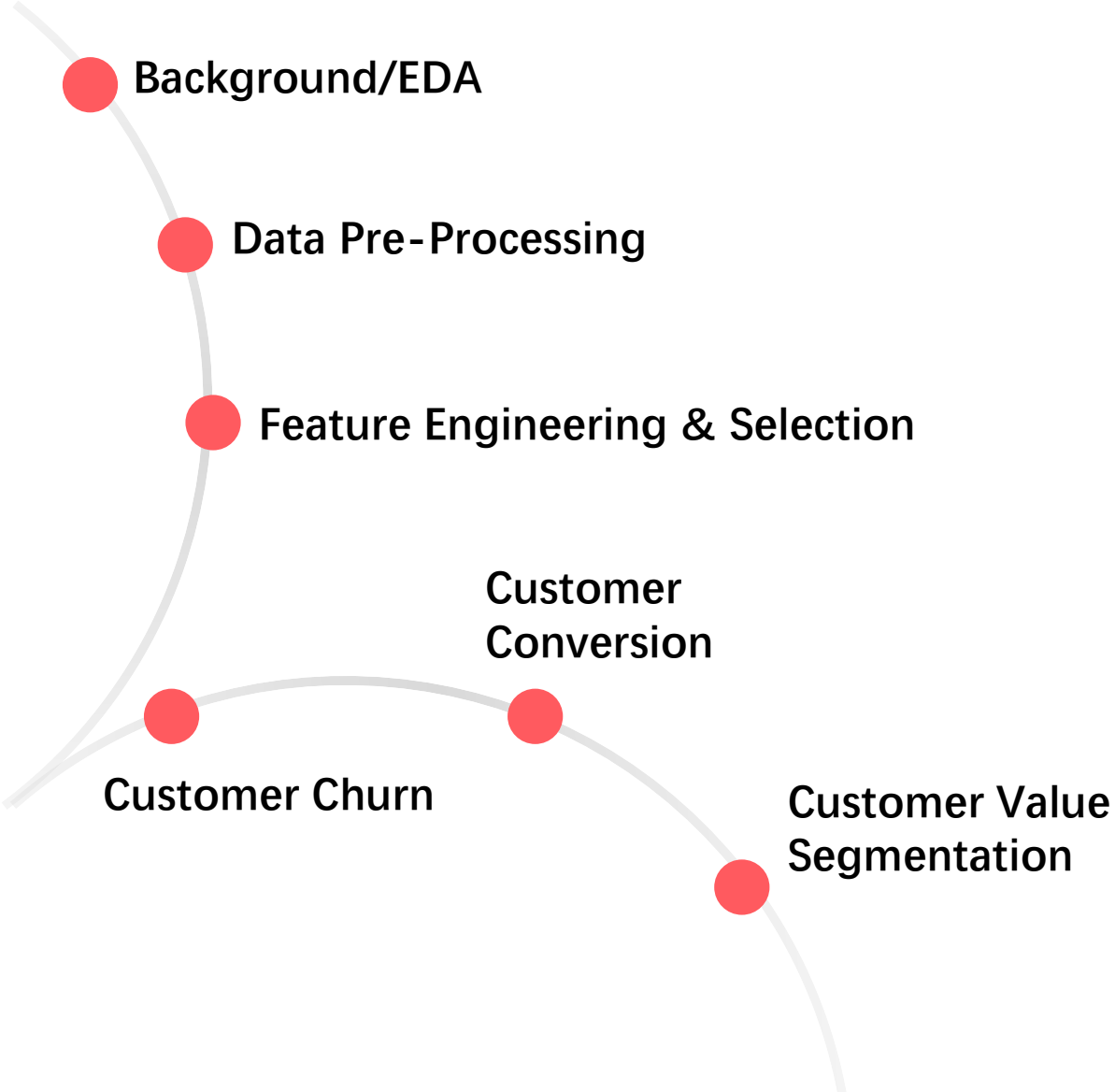
- Project Workflow
- Data Wrangling & Feature Selection
- Random Forest Forecasting - Customer Conversion Analysis
- Logistic Regression - Customer Churn Prediction Analysis
- RFM Modeling (K-Means) - Customer Value Analysis

Airbnb App User Data

Airbnb Open Data –App Users data within one week in the year 2019

Package Used

- Python
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn



Contents

Part1 - Data Wrangling & Feature Selection

Part2 - Customer Conversion Analysis - Random Forest Forecasting

Part3 - Customer Churn Prediction Analysis - Logistic Regression

Part4 - Customer Value Analysis - RFM Modeling (K-Means)

EDA Overview

- 70K Observations
- 50 Features
- Lots of Missing Values
- data sets with high kurtosis/skewness
- No duplicates

Objective

- To conduct Customer Churn Predictive Analysis based on binary label, “churn”.
- To predict Customer Conversion using Random Forest Regression
- To explore Customer Value from RFM Segmentation technique

Background

- The data is produced from Airbnb App, exploring customer behavioral pattern (such as visit time, arrival time, order number, conversion rate, consuming capacity, averaging price, etc.)
- Through data desensitization of user data
- Label variable: 0 represents churn while 1 represents retention

Data Pre-Processing

Missing Values

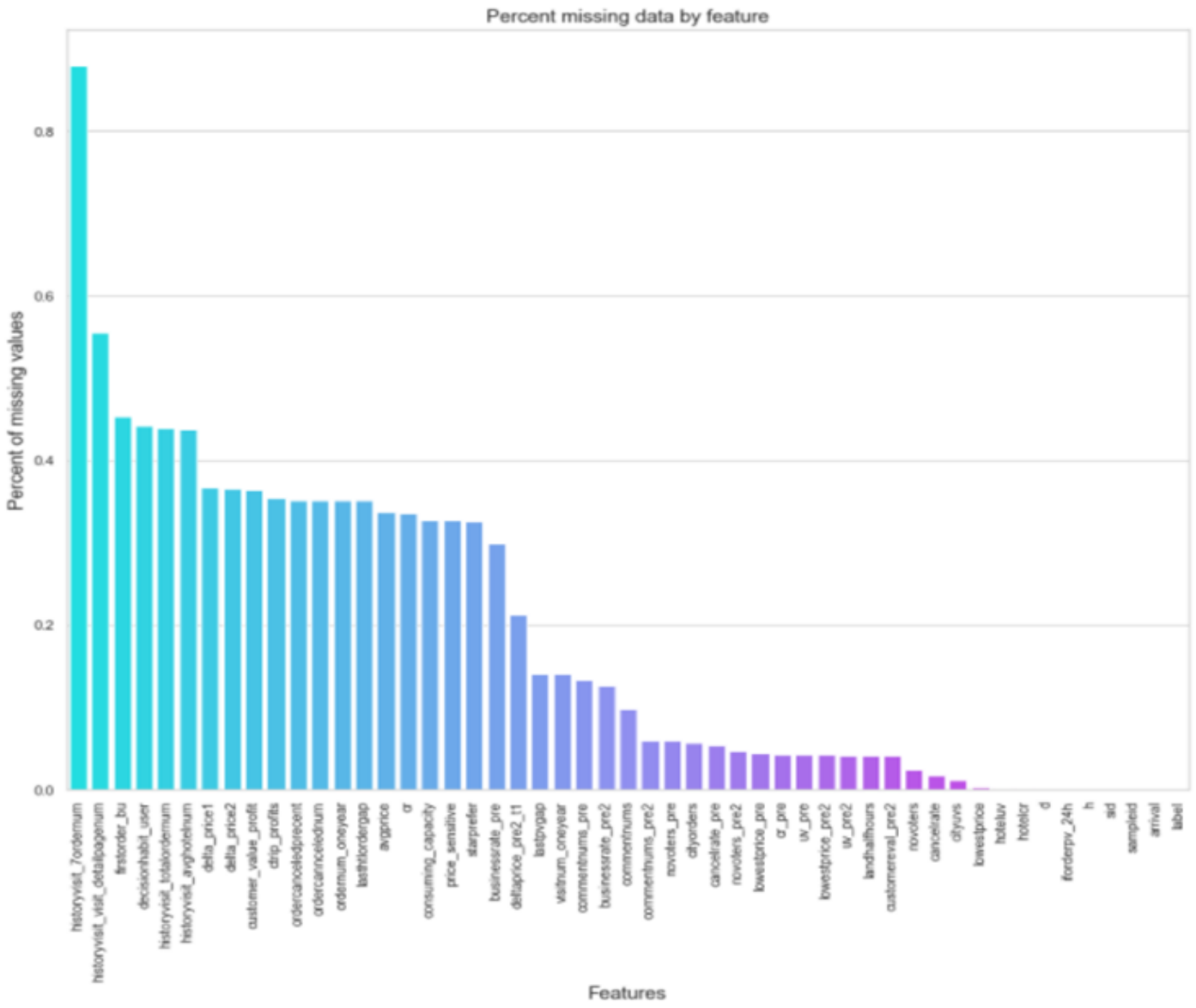
Imputation Rule

- Fill in Median value for features having negative values
- Fill in Mode value for categorical features
- Fill in Mean value for the rest of features
- Remove features with over 70% missing values

Outliers

Capping of outliers using Percentile method

- Any data points less than the value at first percentile or greater than value at 99th percentile could be possible outliers
- Outlier detection and removal using percentile



Feature Selection

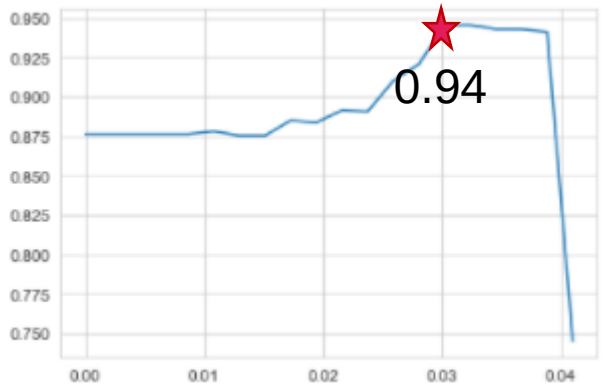
Multiple Feature Selection Methods

Variance Threshold
Feature selector that removes all low-variance features

F-Test
Remove 7 Variables with 0 coefficient

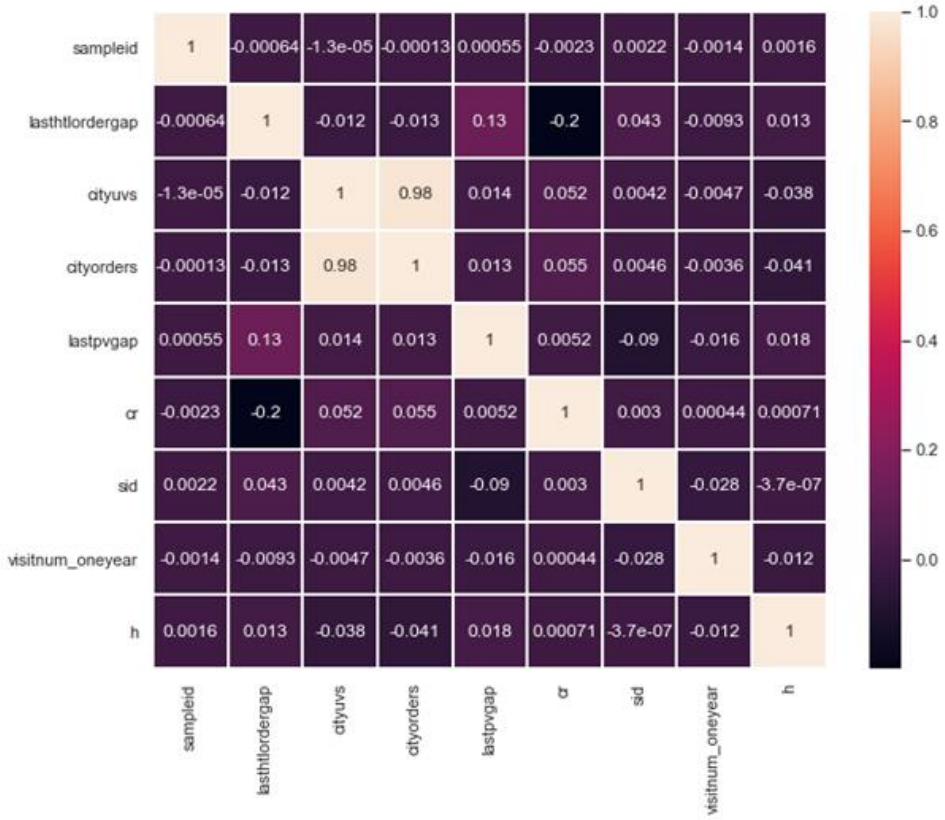
Embedded Methods

- Based on Random Forest Learning Curve, when the threshold equals to .03, the model wins the highest score – .94
- Use the .03 threshold to get 8 features with importance over the threshold:
- 'lasthtlordergap', 'cityorders', ‘cityorders’, 'lastpvgap', 'cr', 'sid','visitnum_oneyear', 'h'



Filter out highly correlated features

It's been identified that there's a high correlation between ‘cityorders’ and ‘cityuvs’.



Feature Engineering

Statistical Data Feature Binning

Selection of Binning Methods | **Data Distribution** | **Equal Width Binning**

Data Distribution

- Except for ‘h’, time of visit, all features data are skewed to right

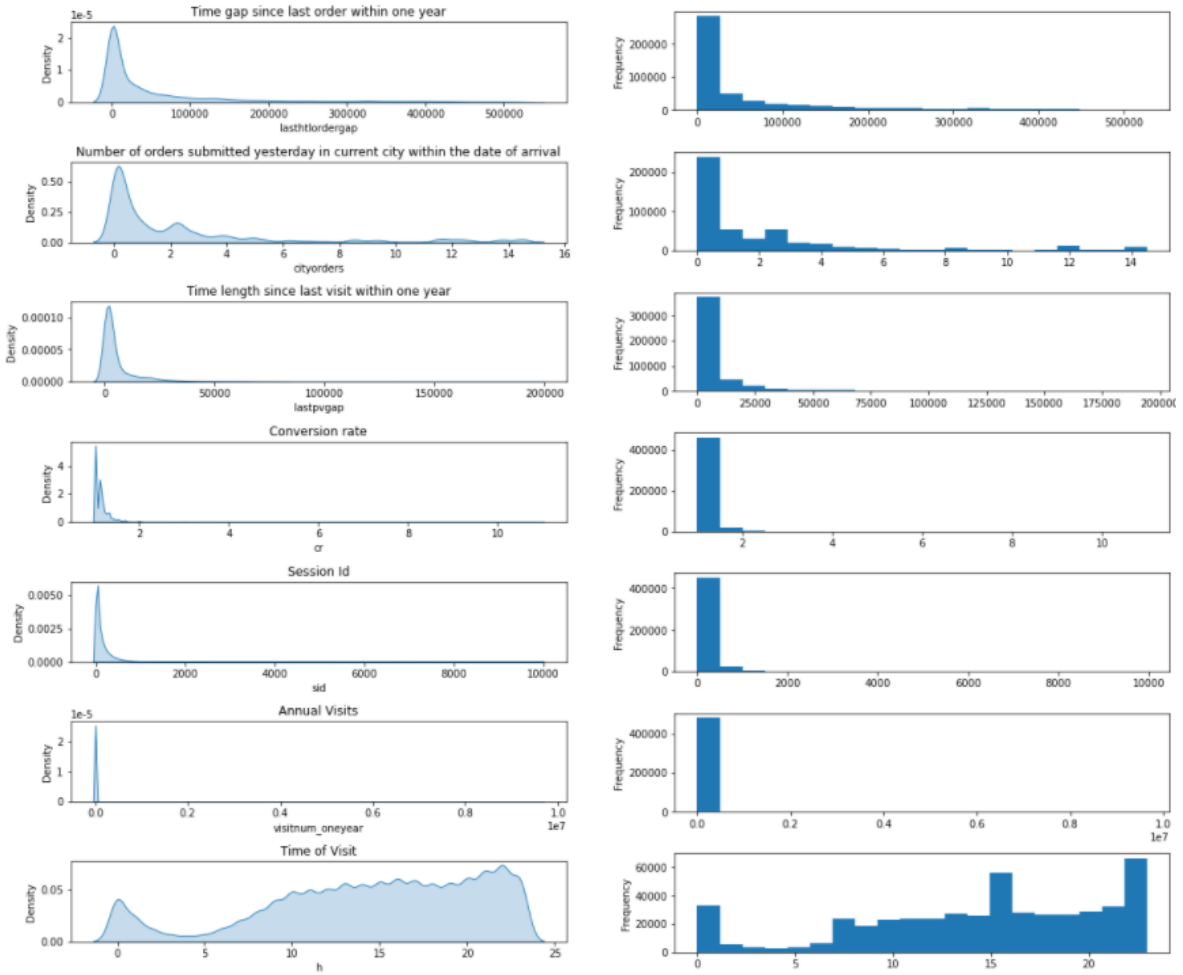
Binning

- Binning is used for the transformation of a continuous or numerical variable into a categorical feature
- Binning of continuous variable introduces non-linearity and tends to improve the performance of the model
- It can be also used to identify missing values or outliers

Unsupervised Binning

- Unsupervised binning transforms variables into categorical bins without considering the target class label into account
- For this project data, it adds limited information as highly skewed data distribution.

- Alternatively, the **entropy-based Supervised Binning** was adopted for bucketing variables.



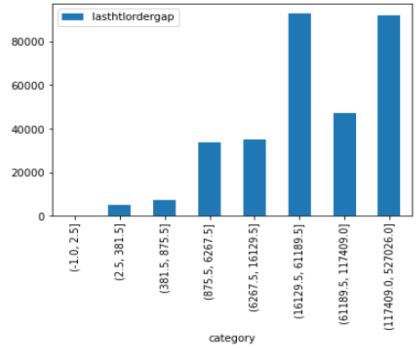
Binning based on CART Tree Impurity Thresholds

Optimal Binning / Supervised discretization

- The algorithm is CART Tree which initially excludes missing values to compute the cut points –
- y is a binary response variable,1 means churn, 0 means retention

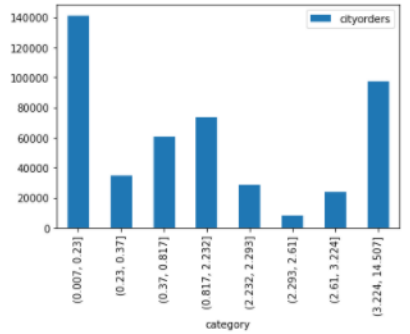
General Steps

1. Use GridSearchCV to adjust parameters and get best leaf's sample percentage
2. Use best parameters to build CART tree model
3. Extract threshold of test node to calculate best cut points/bins



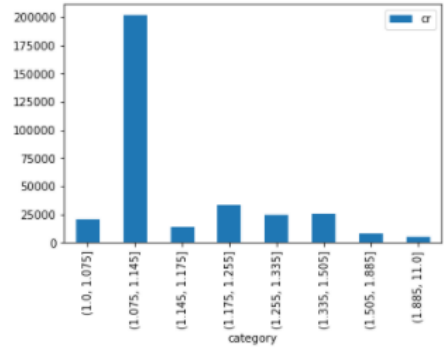
lasthtlordergap

Time gap since last order within one year



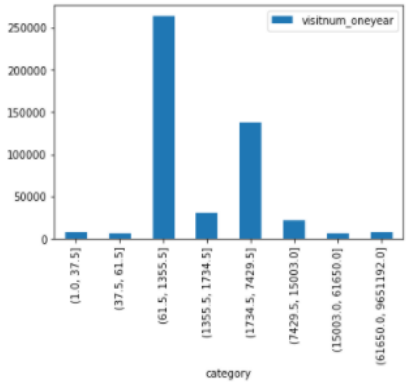
cityorders

Number of orders submitted yesterday in current city within the date of arrival



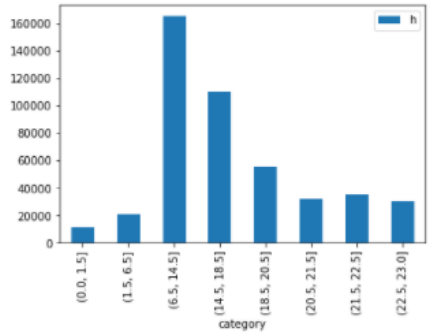
cr

Conversion Rate



visitnum_oneyear

Annual Visit Number



h

Time of Visit

Feature Engineering

Statistical Data Feature Binning

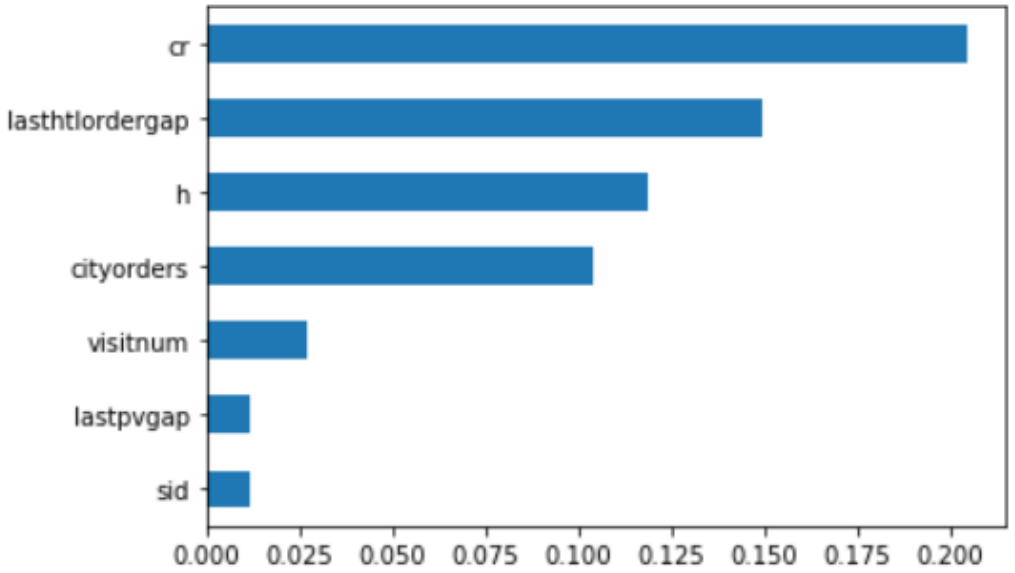
Weight of Evidence Binning

- In logistic regression modeling process, weight of evidence is often used because it handles outliers, missing values, categorical variables appropriately. Also, WoE transformation helps you to build strict linear relationship with log odds.
- The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable
- For this project, the Weight of Evidence for each bin is calculated based on –

$$WOE = Ln\left(\frac{\%\ of\ Customer\ Rentention\ (label\ "1")}{\%\ of\ Customer\ Churn\ (lable\ "0")}\right)$$

$$IV = \sum (\%\ of\ Customer\ Rentention\ (label\ "1") - \%\ of\ Customer\ Churn\ (lable\ "0")) * WOE$$

Information Value

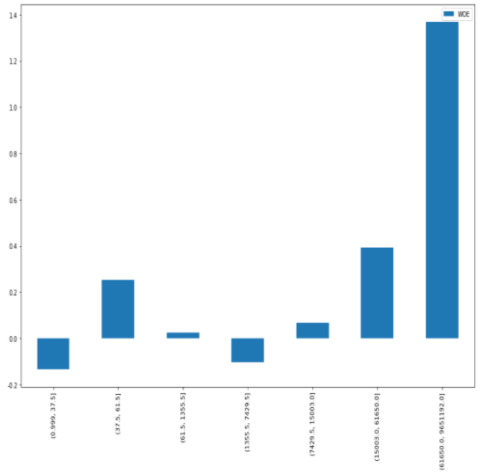


- Information value helps to rank/screen variables on the basis of their importance
- variables with least IV value, ‘lastpvgap’ and ‘sid’, are eliminated

Feature Engineering

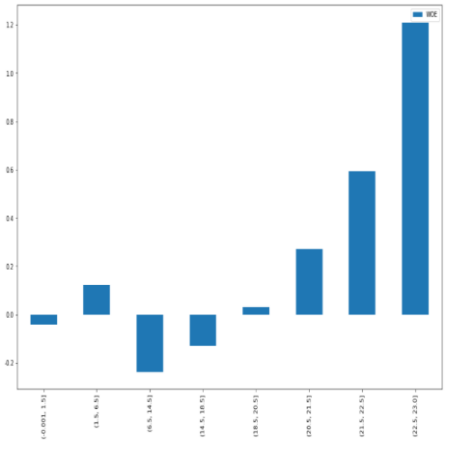
Weight of Evidence Binning

visitnum_oneyear



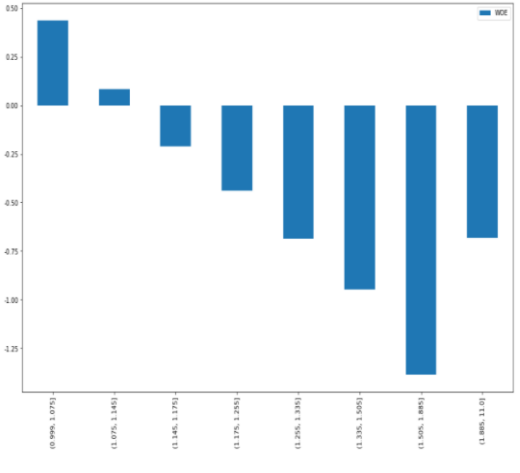
- The retention fluctuates slightly with low number of visits within one year
- Only when the number of visits reaches a certain level (more than 15,000), the app gains significant customer retention and peaks after 61,650

h



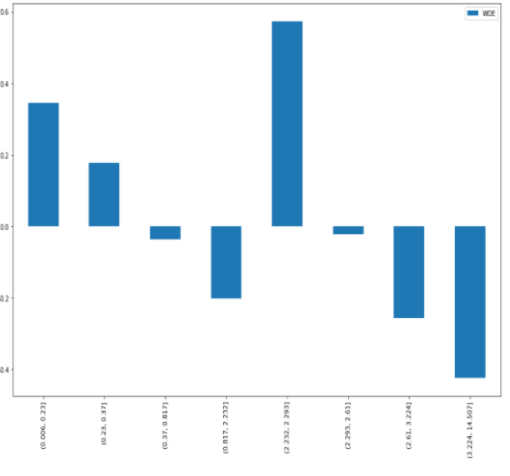
- The app has no good retention until 6:30 PM
- In the evening, the customer retention ramps up and peaks between 10:30 PM and 11 PM.

cr



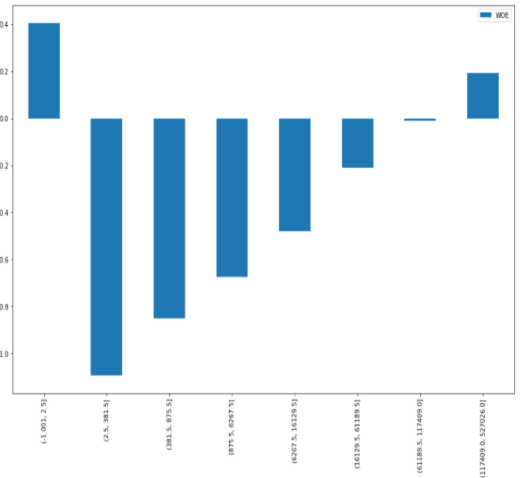
- When conversion rate is less than 1.14%, users retain relatively more users.
- As conversion rate increases, churn rate also increases, but it should be noted that when conversion rate is over 1.89%, customer churn slowed down.

cityorders

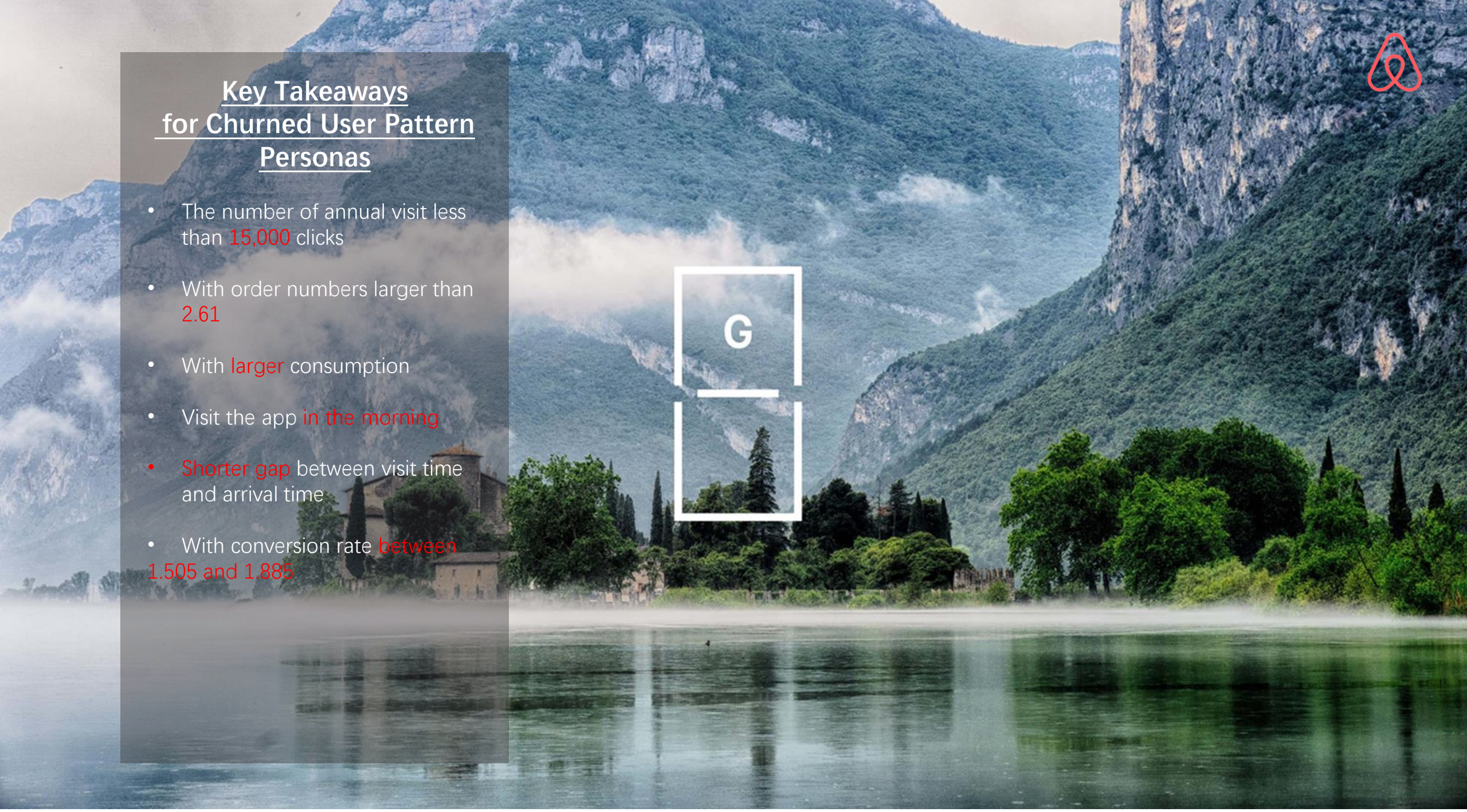


- For cityorders in the 5th box (2.23, 2.293) interval, the app has best retention.
- Except for box 5, the results of customer retention gradually decrease as the number of orders submitted yesterday increases.

lasthtlordergap



- For the lasthtlordergap interval of (2.5, 381.5), the app has most severe churn.
- As time goes by, the risk of churn decreases.



Key Takeaways for Churned User Pattern Personas

- The number of annual visit less than 15,000 clicks
- With order numbers larger than 2.61
- With larger consumption
- Visit the app in the morning
- Shorter gap between visit time and arrival time
- With conversion rate between 1.505 and 1.885

Contents

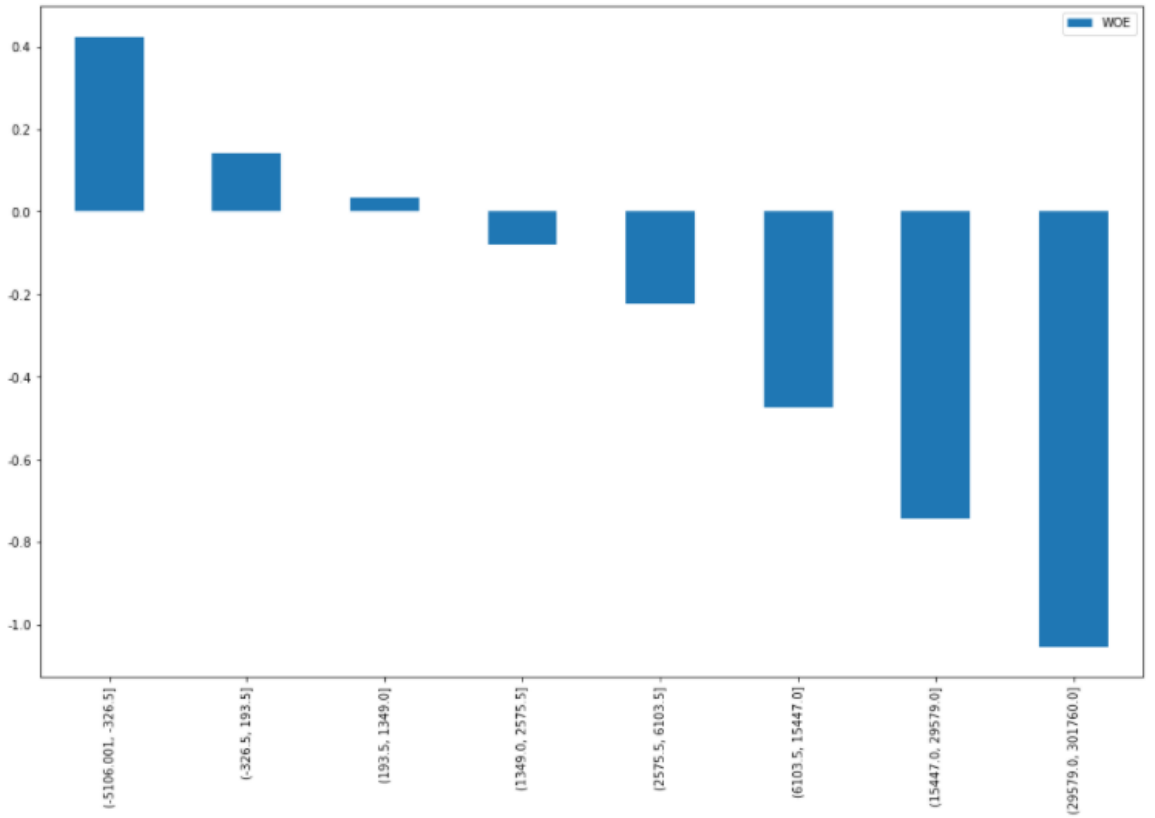
Part1 - Data Wrangling & Feature Selection
Part2 – Logistic Regression - Customer Churn Prediction Analysis
Part3 - Random Forest Forecasting - Customer Conversion Analysis
Part4 - RFM Modeling (K-Means) - Customer Value Analysis

Logistic Regression

Creation of New Feature

Monetization of Order Consumption

- Creation of new feature, ‘M’, Consumption monetary value
- M = ordernum_oneyear * avgprice.
- The ‘M’ has relatively higher IV. (0.12)
- We multiply the average price and the annual order volume to get the annual consumption, find out that the more customers consume each year, the more churn exists.
- The overall trend is diminishing. **The more customers spend each year, the greater the risk of attrition.**

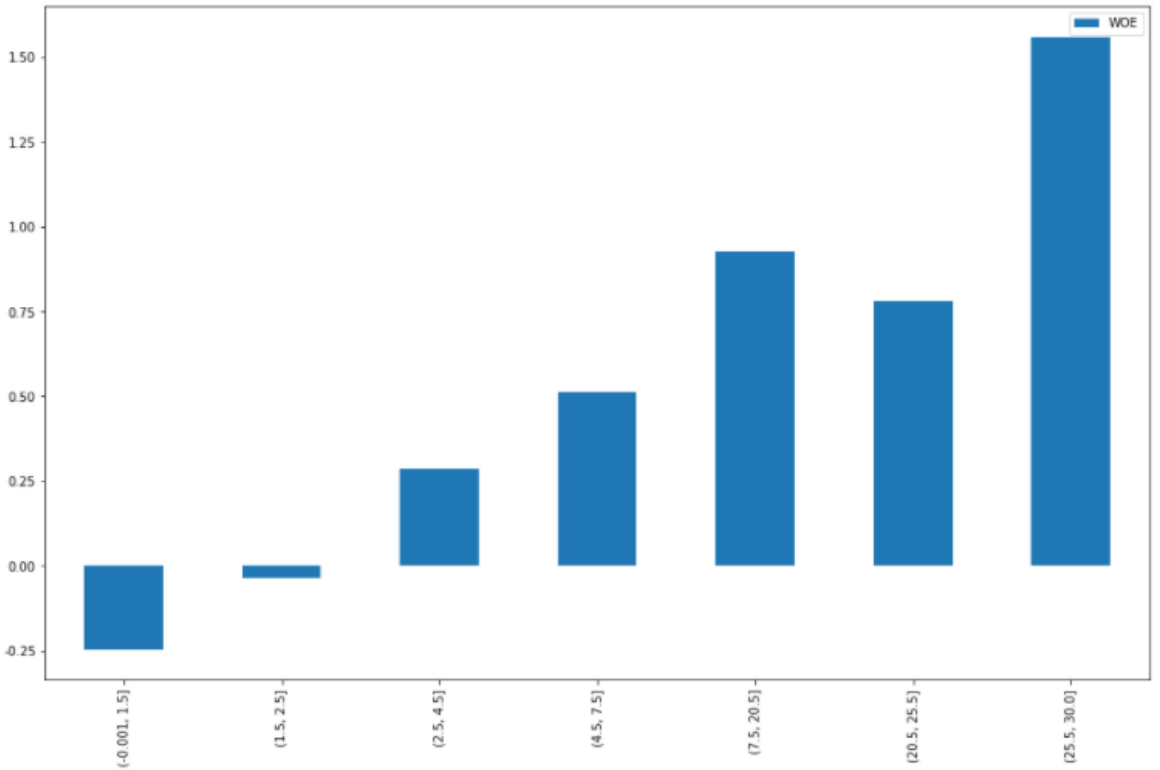


Logistic Regression

Creation of New Feature

Delta – Time Difference

- Creation of new feature, ‘Delta’, time difference between app visit date and arrival date
- Delta = arrival time – visit time.
- The ‘Delta’ has relatively higher IV. (0.17)
- From the perspective of bins, **the longer the gap between visit date and arrival date, the less the churn is**



Logistic Regression

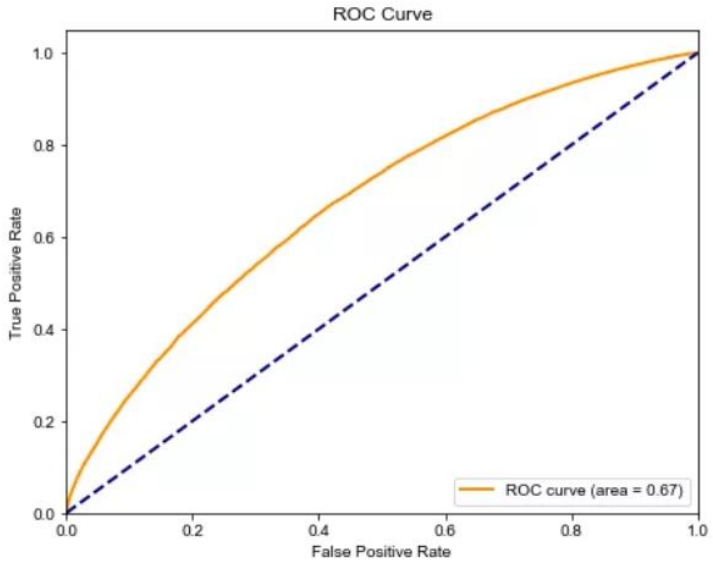
Before adding 2 new features

auc	accuracy	precision	recall	f1
0.669	0.736	0.750	0.945	0.835

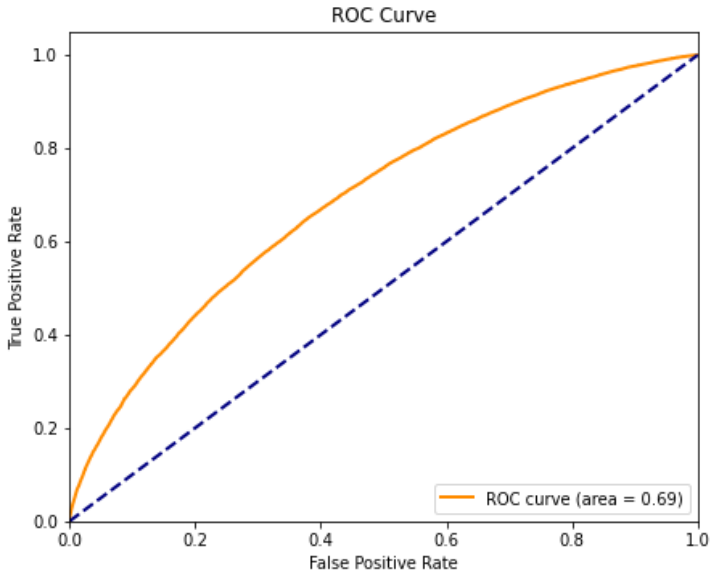
After adding 2 new features

auc	accuracy	precision	recall	f1
0.690	0.737	0.751	0.953	0.840

- ROC Score increased 2%
- Logistic Regression Score is 0.737



Before adding 2 new features, AUC = 0.67



After adding 2 new features, AUC = 0.69

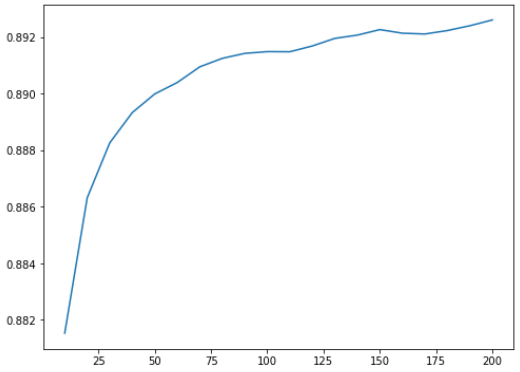
Contents

Part1 - Data Wrangling & Feature Selection
Part2 - Logistic Regression - Customer Churn Prediction Analysis
Part3 - Random Forest Forecasting - Customer Conversion Analysis
Part4 - RFM Modeling (K-Means) - Customer Value Analysis

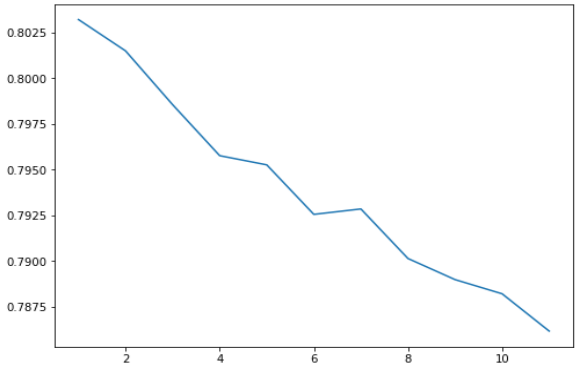
Random Forest Forecasting Model

Hyper Paremeter Tuning

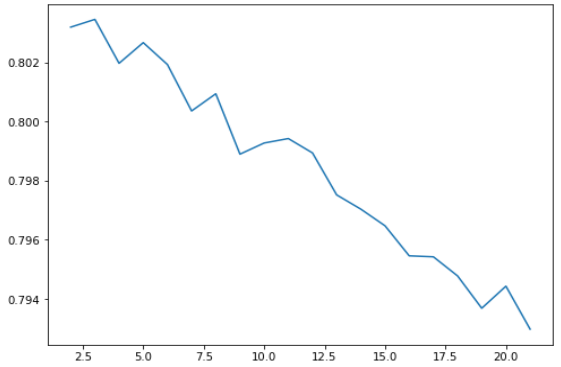
- Random Forest is an advanced ensemble technique in predicting customer churn
- Use grid search to adjust parameters for below Optimum parameter portfolio -



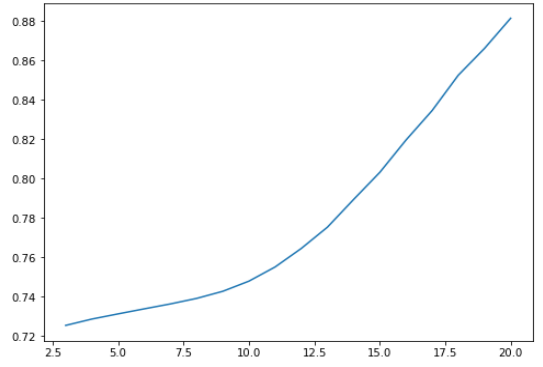
N estimators=200



Min samples leaf=1



Min samples split=3



Maximum Depth = 20

Random Forest Customer Churn Analysis

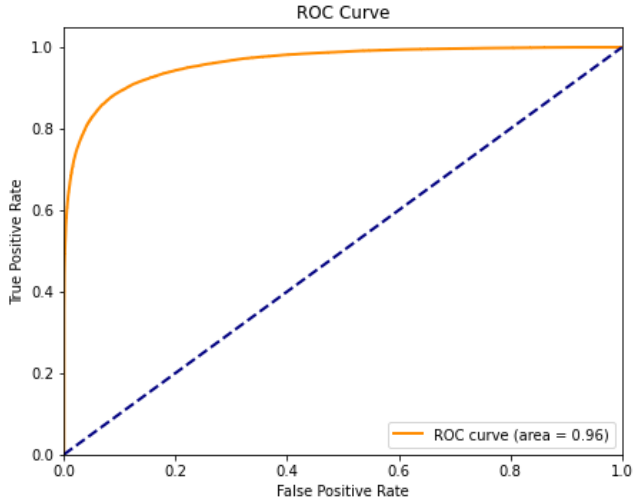
Best Parameters

n_estimators=200
max_depth=20
min_samples_leaf=1
min_samples_split=3

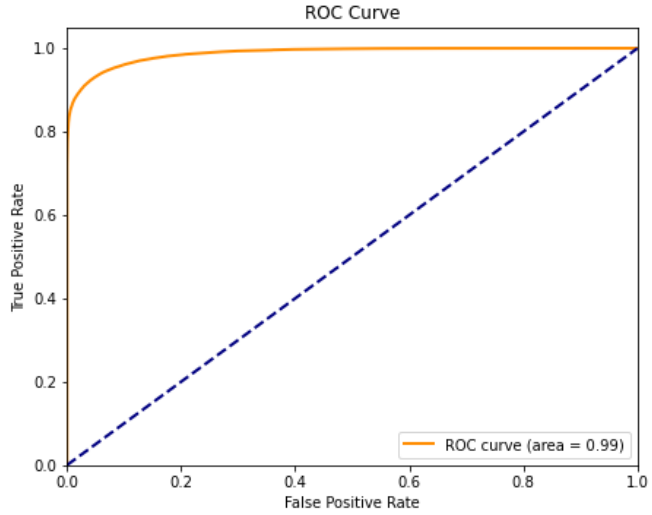
Evaluation Results

Finally, the random forest classifier ensures that the ROC curve area reaches **0.97** without overfitting.

Before Hyper parameter Tuning



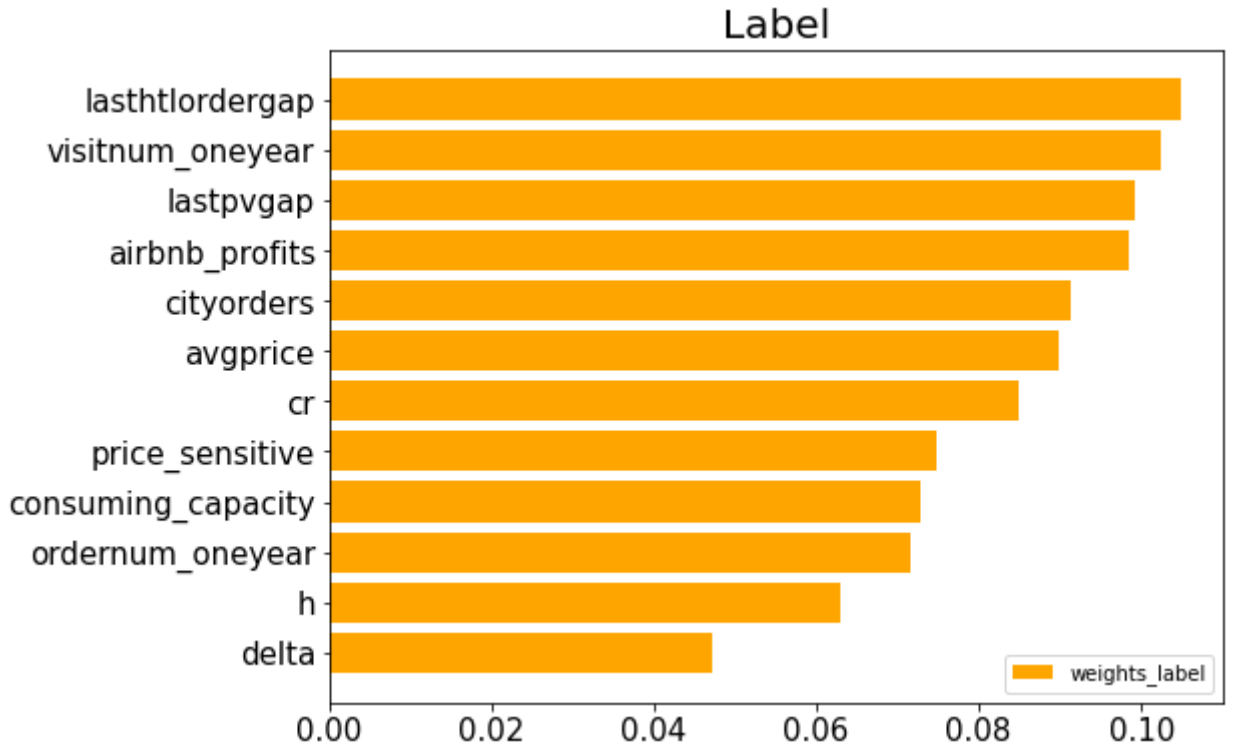
After Hyper parameter Tuning



Random Forest Customer Churn Analysis

The Top 5 Importance Features

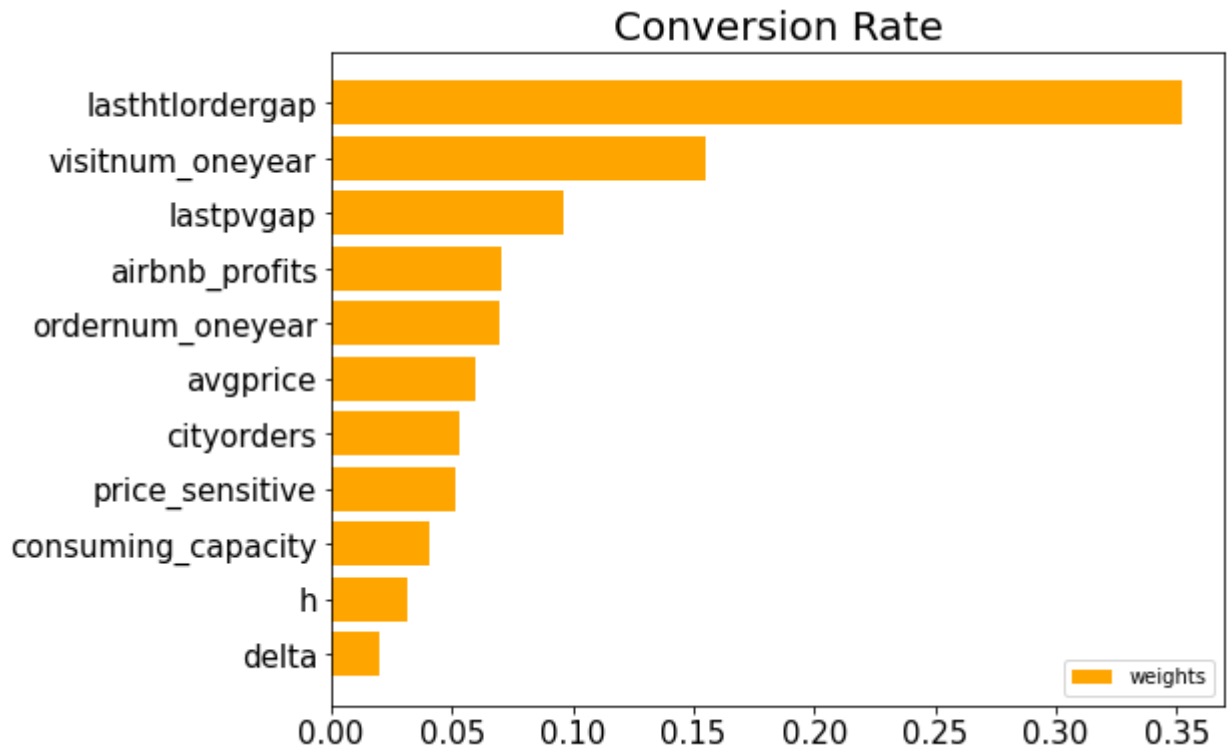
- Lasthlordergap
- Visitnum_oneyear
- Lastpvgap
- Airbnb_profits
- Cityorders



Random Forest Customer Conversion Analysis

The Top 5 Importance Features

- Lasthlordergap
- Visitnum_oneyear
- Lastpvgap
- Airbnb_profits
- Ordernum_oneyear



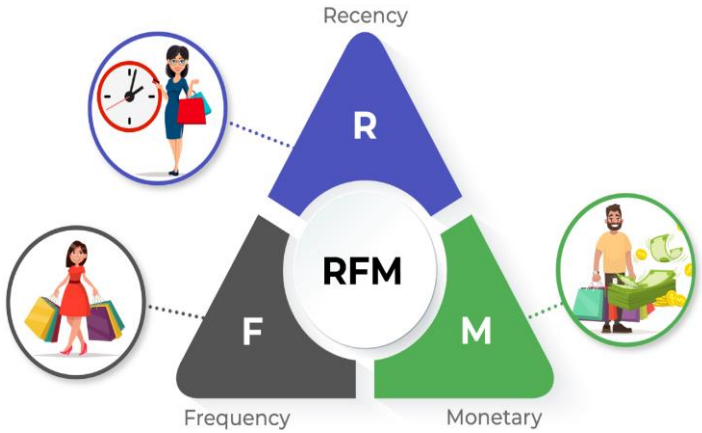
Contents

Part1 - Data Wrangling & Feature Selection
Part2 - Random Forest Forecasting - Customer Conversion Analysis
Part3 - Logistic Regression - Customer Churn Prediction Analysis
Part4 - RFM Modeling (K-Means) - Customer Value Analysis

Introduction to RFM Model

A data-driven customer behavior segmentation technique

1. Recency, frequency, monetary value (RFM) is a marketing analysis tool used to identify a firm's best clients based on the nature of their spending habits.
2. An RFM analysis evaluates clients and customers by scoring them in three categories: how recently they've made a purchase, how often they buy, and the size of their purchases.
3. RFM analysis helps firms reasonably predict which customers are likely to purchase their products again, how much revenue comes from new (versus repeat) clients, and how to turn occasional buyers into habitual ones.



R(Recency)

- Time difference between the customer's last transaction
- The larger the R value, the longer the date the customer transaction occurred.
- User stickiness, the smaller the better

F (Frequency)

- The number of transactions by the customer in the most recent period of time
- The larger the F value, the more frequent customer transactions
- User loyalty, the bigger the better

M(Monetary)

- The amount of transactions made by the customer in the most recent period of time.
- The larger the M value, the higher the customer value.
- User contribution, the bigger the better

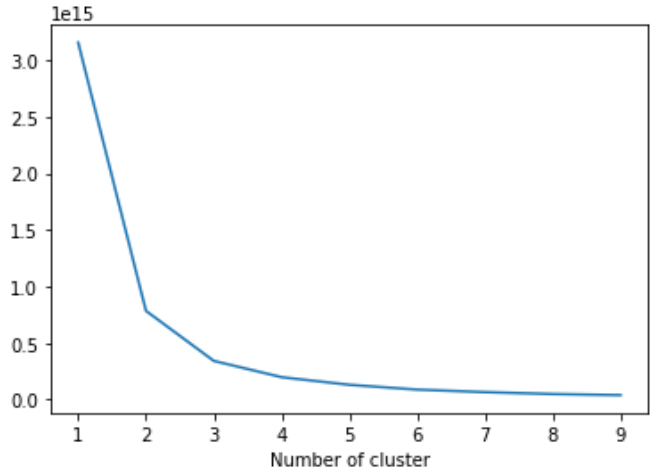
RFM K-means Clustering

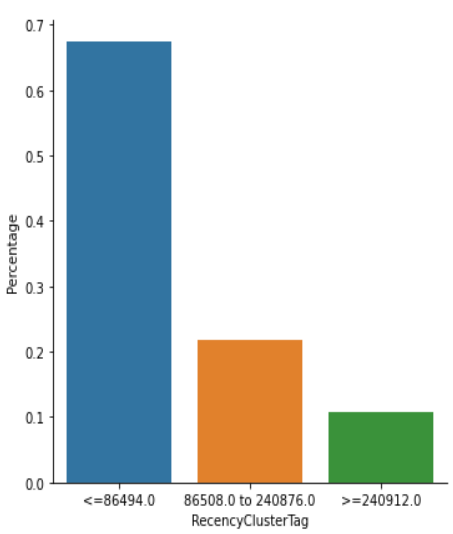
Selected Features

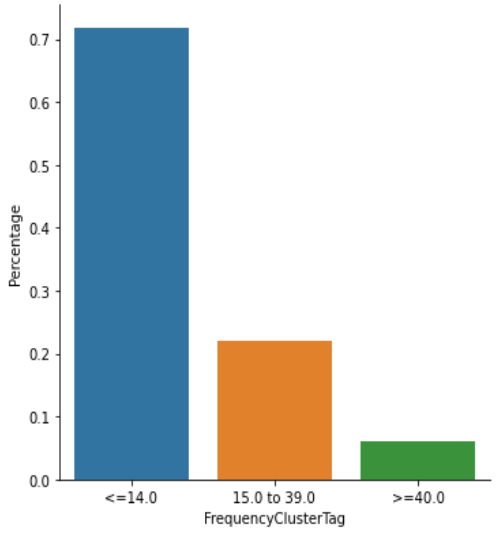
RFM	R	F	M
Model	lasthtlorderg ap	ordernum_o neyear	avgprice

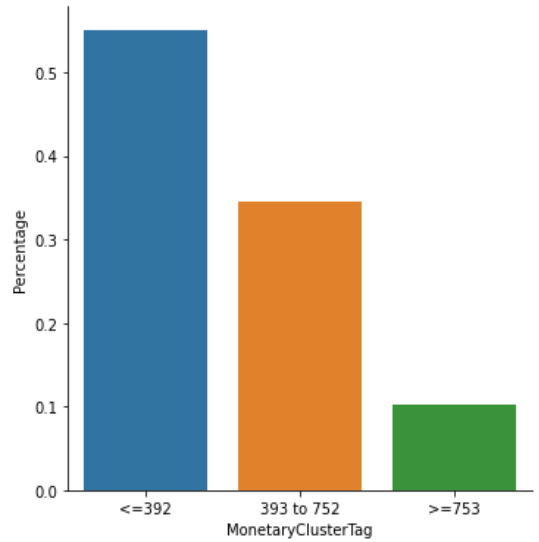
Unsupervised Learning

- Using K means to determine RFM Clusters
- In accordance with Elbow Method, the best clusters are 3.









Label K-mean Cluster for RFM

- Encode the cluster by 1-3 order
- 1 represents low score (worse)
- 2 represents median score (fair)
- 3 represents higher score (better)
- For recency, the higher recency, the lower score

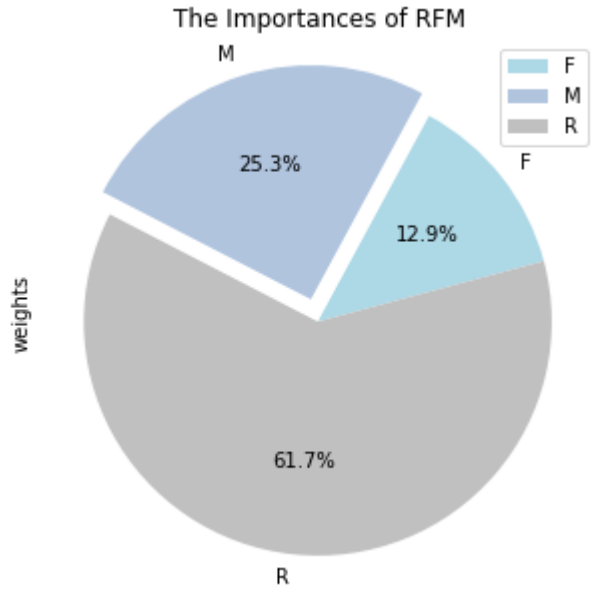
Selected Feature Importance

- Use RandomForest (Retention) to determine feature importance for RFM, respectgively

RFM	R	F	M
Model	lasthtlorderg ap	ordernum_o neyear	avgprice
Weight	0.6174	0.1294	0.2531

Random Forest Modeling

	sampleid	F	M	R	R_score	F_score	M_score
0	24650	21.0	363.0	10475.0	3	2	1
1	24653	7.0	307.0	18873.0	3	1	1
2	24658	33.0	1000.0	4616.0	3	2	3
3	24662	4.0	685.0	44830.0	3	1	2
4	24665	7.0	407.0	5823.0	3	1	2
...	...	...	...	...	...	...	...
288042	2238388	2.0	226.0	170680.0	2	1	1
288043	2238389	4.0	461.0	528.0	3	1	2
288044	2238396	5.0	193.0	63673.0	3	1	1
288045	2238397	1.0	258.0	125643.0	2	1	1
288046	2238403	3.0	256.0	75105.0	3	1	1



Overall Customer Value Score

0.6174*R score
+
0.1294*F score
+
0.2531*M score

Introduction to RFM Model

A data-driven customer behavior segmentation technique

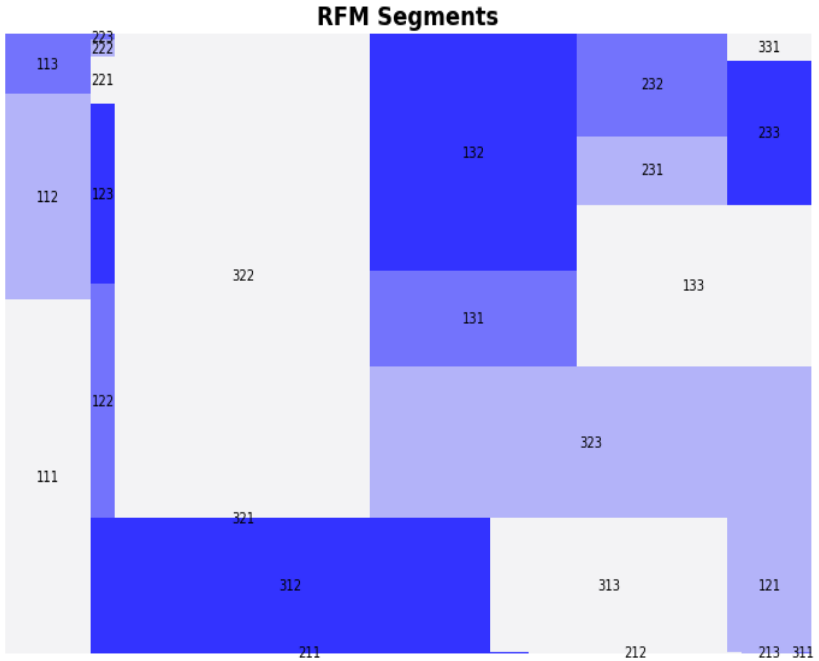
- Create RFM Segments based on K-means clusters
- Below are 5 largest groups from Airbnb App Users –

RFM Group	Number
311	71,187
312	38,478
211	30,929
321	28,345
322	21,795

- General speaking, the app still has lots of new users who haven't engaged in the use of app for long time

- Therefore, customer success team need to focus on improving user experience of new customer group

R	F	M	Segmentation
1	1	1	churned customers
1	1	2	churned customers
1	1	3	Important customer for retention
1	2	1	General customers for maintenance
1	2	2	General customers for maintenance
1	2	3	General customers for maintenance
1	3	1	General customers for maintenance
1	3	2	General customers for maintenance
1	3	3	Important customer for recall
2	1	1	General customers for maintenance
2	1	2	General customers for maintenance
2	1	3	General customers for maintenance
2	2	1	General customers for maintenance
2	2	2	General customers for maintenance
2	2	3	Important customer for retention
2	3	1	General customers for maintenance
2	3	2	General customers for maintenance
2	3	3	Important customer for recall
3	1	1	New Customer
3	1	2	General customers for maintenance
3	1	3	Important customers for development
3	2	1	General customers for maintenance
3	2	2	New Customer
3	2	3	Important customers for development
3	3	1	Customers with potentials
3	3	2	Customers with potentials
3	3	3	Core Valuable Customers



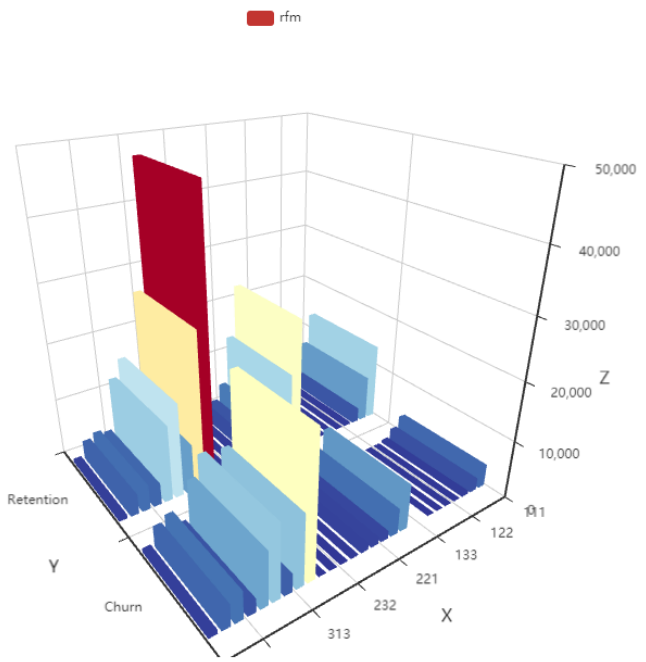
RFM Modeling

By Number of Users

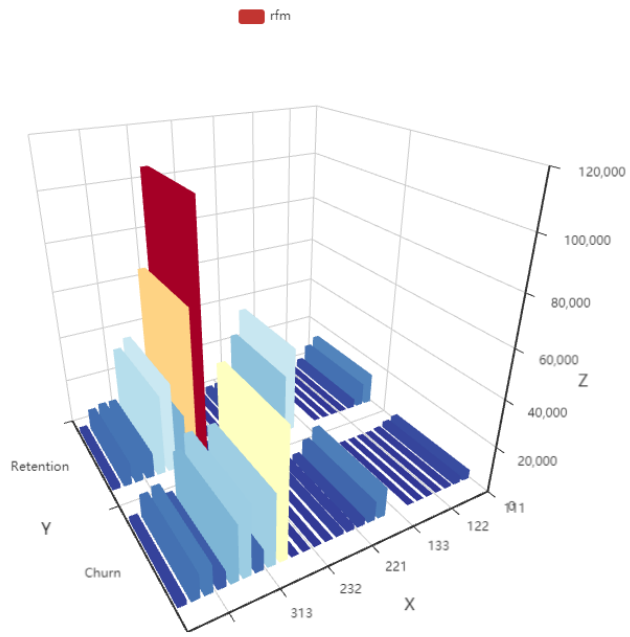
- Segment Customers based on RFM Groups
- Height is measured by number of users
- Customer Segments with best retention –
- 133, 233, 311
- Customer Segments with lots of churn –
- 233, 311, 313

By Total Adjusted Values of Users

- Segment Customers based on RFM Groups
- Height is measured by total adjusted values of users
- Customer Segments with most retention values –
- 133, 233, 313
- Customer Segments with most of churn values –
- 133, 233, 313



By Number of Users



By Total Adjusted Values of Users

RFM Modeling

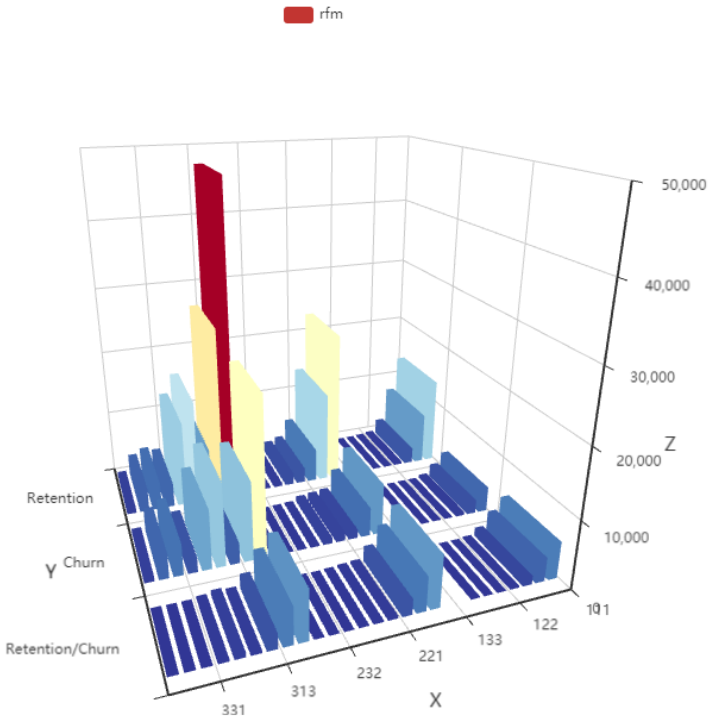
Adding new measurement

- Retention/Churn represents the difference between number of user retention and churn quantify logarithm of the odds ratio for customer value score –
- Formula: (number of retention – number of churn)* log(sum(rfm score) of retention/sum(rfm score) of churn)
- This new measurement can be used to quantify the real value of each segments

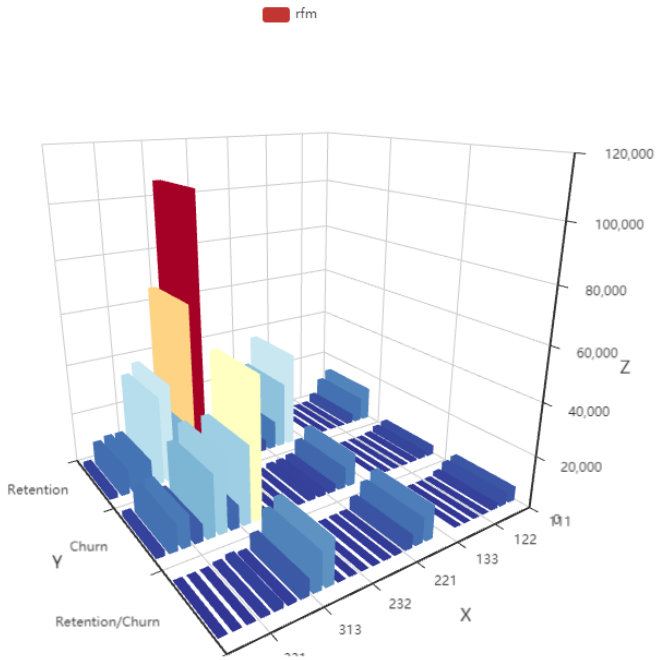
Key Takeaways

- Based on RFM Modeling & Analysis, below 3 segments of Airbnb App users take the majority of real market value
- Airbnb need to focus on customer win of recalling customers and attracting new customers with Users having long recency or new customer most recently visit the app

R	F	M	Segmentation
1	3	3	Important customer for recall
2	3	3	Important customer for recall
3	1	1	New Customer



By Number of Users



By Total Adjusted Values of Users

THANK YOU