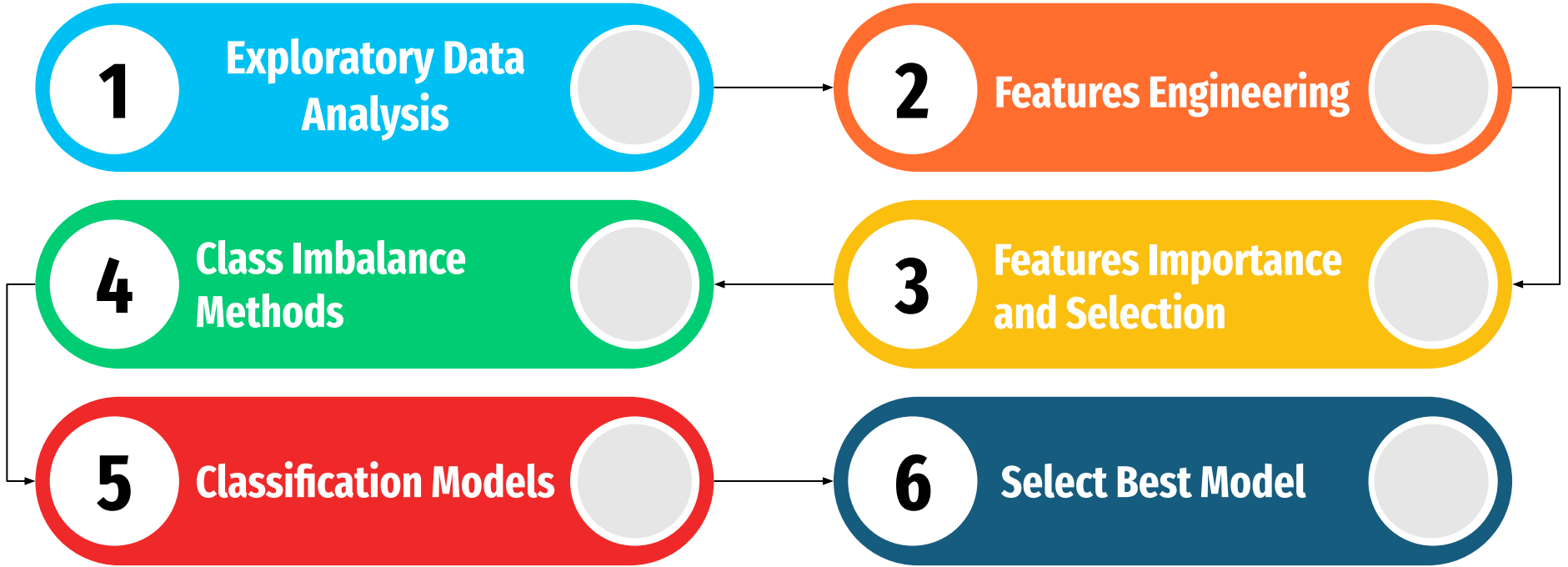
Online Shopping Products purchase prediction

Raed Altuwaijri

Problem statement

We are an online shopping store our goal is to increase sales and reduce logistics time and costs by recommending a products to the customer base on there prior orders

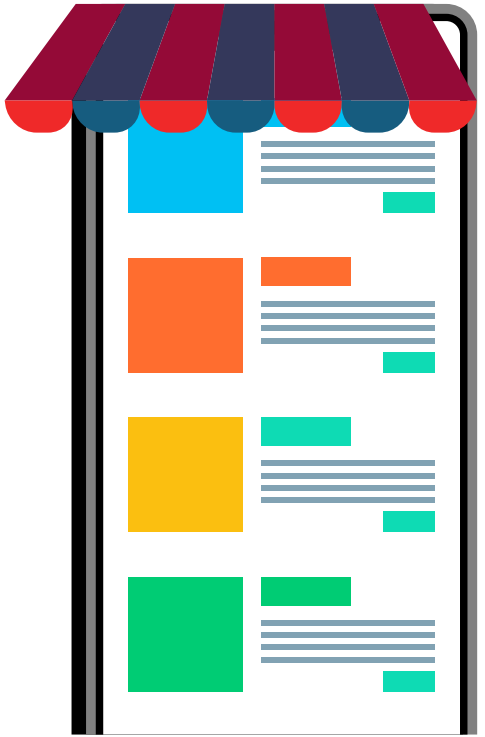
Methodology



Data Description

Data sets

Data about each order and the user places that order	Customer Orders
Data about each product	Products
Data about products in each cart	Cart Products



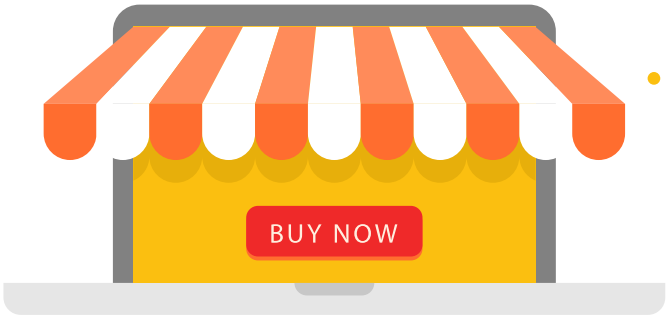
Exploratory Data Analysis

+ 200 Thousand Customer

Saturday evenings and Sunday mornings are the prime time for orders

Every Customer have between 4 t0 100 Orders

+ 3 Million Order



Fruits and Vegetables are the Highest Ordered Products

Features

In_cart: is product in cart? **(the target)**

Order_id: unique identifier for each order

Product_id: unique identifier for each product

Add_to_cart_order: the sequence in which they have been added to the cart in that order.

Reordered: customer reorders count for this product

User_id: unique identifier for each user

Order_number: user order number

Order_dow: order day of week

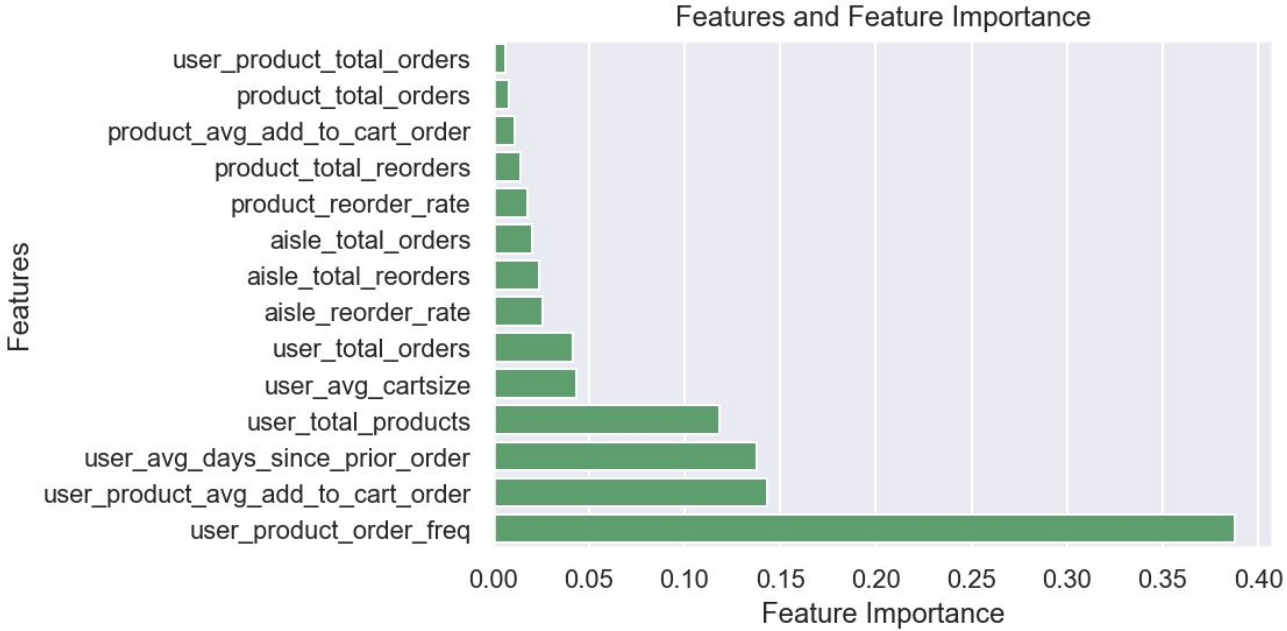
Order_hour_of_day: order hour of day

Days_since_prior_order: days between this order and the prior.

Feature Engineering

User_product_order_freq
Product_reorder_rate
User_total_products
User_avg_days_since_prior_order
User_avg_cartsize
User_product_total_orders
Product_avg_add_to_cart_order
User_product_avg_add_to_cart_order
Product_total_orders
Product_total_reorders
User_total_orders
Aisle_reorder_rate
Aisle_total_reorders
Aisle_total_orders

Feature Importance



Class Imbalance Methods

Adjusting Threshold

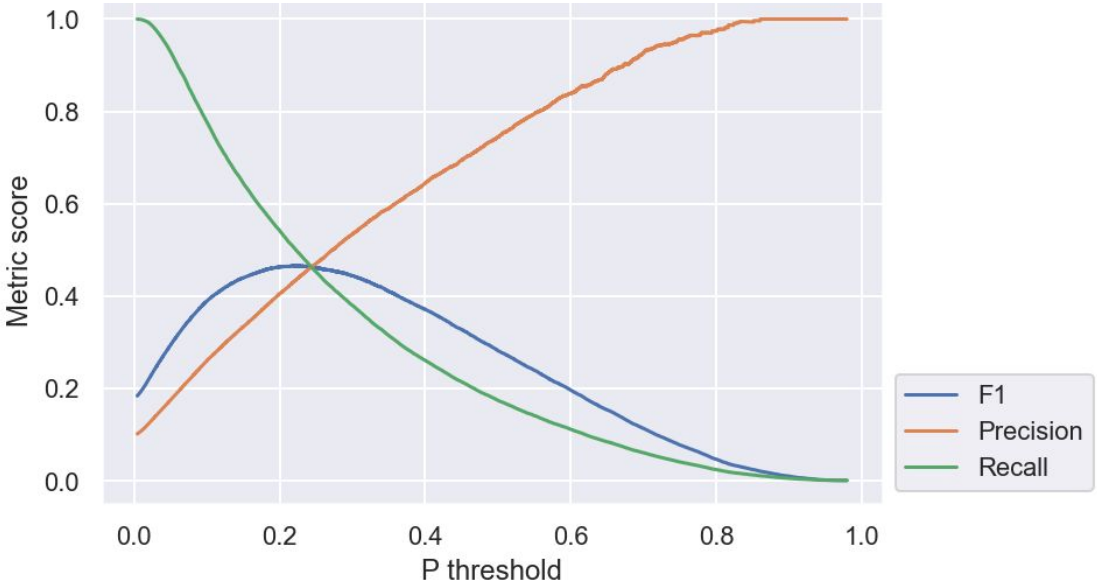
SMOTE Over Sample

Random Under Sample

Random Under Sample

Class Imbalance - Adjusting Threshold

XGB Metric Scores vs Probability Threshold



Best Threshold : 0.22

Best Train F1 Score : 0.465

Classification Models

1	2	3	4	5
Logistic Regression	KNN	Decision Tree	Random Forest	XGB

Model Benchmarking

Scores Based on Competition Leaderboards

Rank	F1 Score
1ST	0.40914
2ND	0.40820
3RD	0.40810
4TH	0.40744

Included in The Prize

Not Included in The Prize

F1 Scores (Testing)

Model	Adjust threshold	SMOTE	Random Over Sample	Random Under Sample
LR	0.327740	0.317308	0.318361	0.294702
DT	0.223622	0.223012	0.221520	0.219156
RF	0.239094	0.345359	0.256328	0.350634
KNN	0.085911	0.217676	0.214491	0.218282
XGB	0.370872	0.334560	0.337023	0.329331

Best F1 Scores (Testing)

Model	Adjust threshold	SMOTE	Random Over Sample	Random Under Sample
LR	0.327740	0.317308	0.318361	0.294702
DT	0.223622	0.223012	0.221520	0.219156
RF	0.239094	0.345359	0.256328	0.350634
KNN	0.085911	0.217676	0.214491	0.218282
XGB	0.370872	0.334560	0.337023	0.329331

Future Work

Use Multiprocessing to Handle Big Data

Cross Validation Based on Time

Hyperparameter tuning