Capstone Final Report

--Repeated Buyers Forecasting

Problem Statement

Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Black Friday or China's Double 11)" in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long-lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI).

For this capstone project, data sets of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day were obtained and the probability that these new buyers would purchase items from the same merchants again within 6 months will be predicted based on the models built. Nevertheless, predictions of the numbers and/or the values of the repeated purchasing are not the goal of this analysis.

Data Wrangling

The raw data set are downloaded from Alibaba website, contained 3 tabular files about user result, user information, and user activity. For instance, I grouped the user information and activity data by user id and merchant id, and then joined all 3 files together. Finally, I got 1 tabular dataset with all necessary features, including user_id – a unique id for the customer

merchant_id – a unique id for the merchant

label – target feature, whether customer is repeated buyer

age_range – age range of customer

gender – gender of customer

total_logs – activity counts of customer

diff_item_review – different items reviewed by customer

diff_cat_review – different item categories reviewed by customer

browse_days – days since first browse record occurred

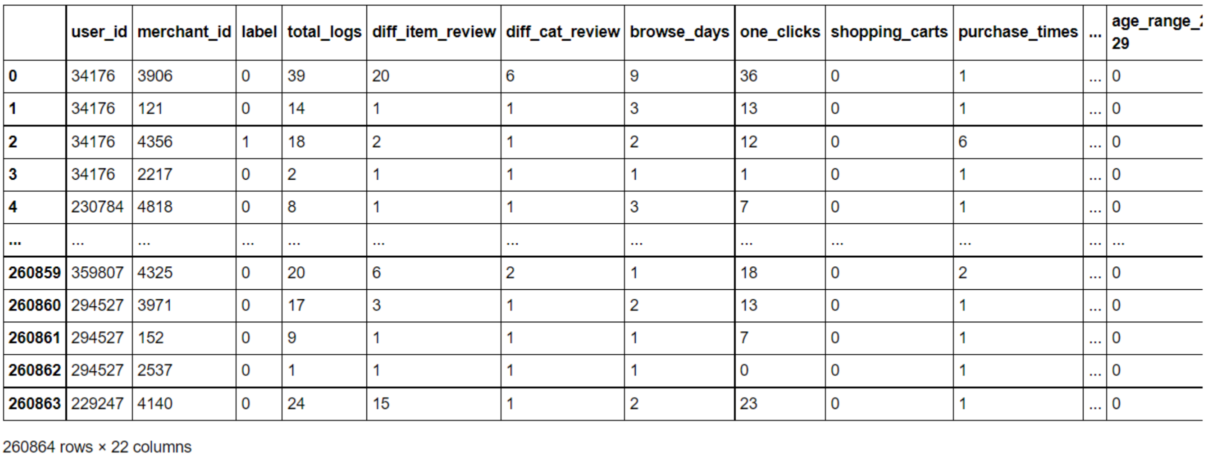
one_clicks – counts of click-only on item

shopping_carts – counts of add-to-cart-only on item

purchase_times – counts of purchase

favourite_times – counts of add-to-favourite-only on item

After dealing with missing values and one hot encoding for categorical features, such as age_range and gender, the dataset contains 260,864 rows of record and 22 columns of features.

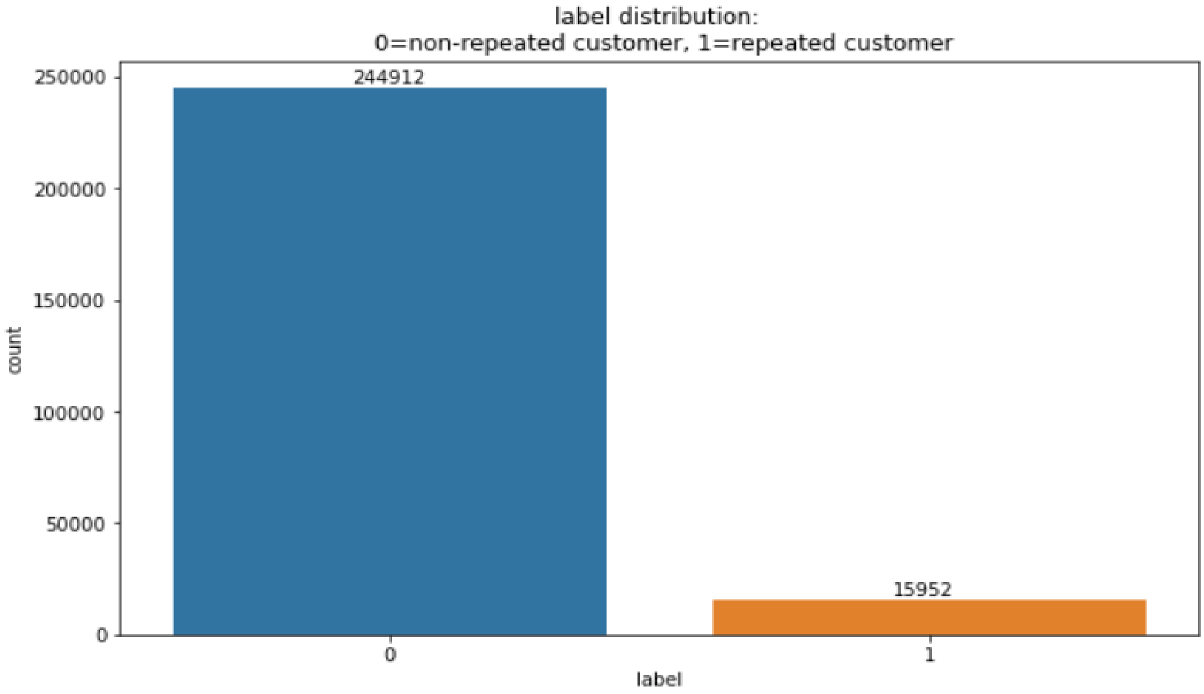


	user_id	merchant_id	label	total_logs	diff_item_review	diff_cat_review	browse_days	one_clicks	shopping_carts	purchase_times	...	age_range_ 29
0	34176	3906	0	39	20	6	9	36	0	1	...	0
1	34176	121	0	14	1	1	3	13	0	1	...	0
2	34176	4356	1	18	2	1	2	12	0	6	...	0
3	34176	2217	0	2	1	1	1	1	0	1	...	0
4	230784	4818	0	8	1	1	3	7	0	1	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
260859	359807	4325	0	20	6	2	1	18	0	2	...	0
260860	294527	3971	0	17	3	1	2	13	0	1	...	0
260861	294527	152	0	9	1	1	1	7	0	1	...	0
260862	294527	2537	0	1	1	1	1	0	0	1	...	0
260863	229247	4140	0	24	15	1	2	23	0	1	...	0

260864 rows × 22 columns

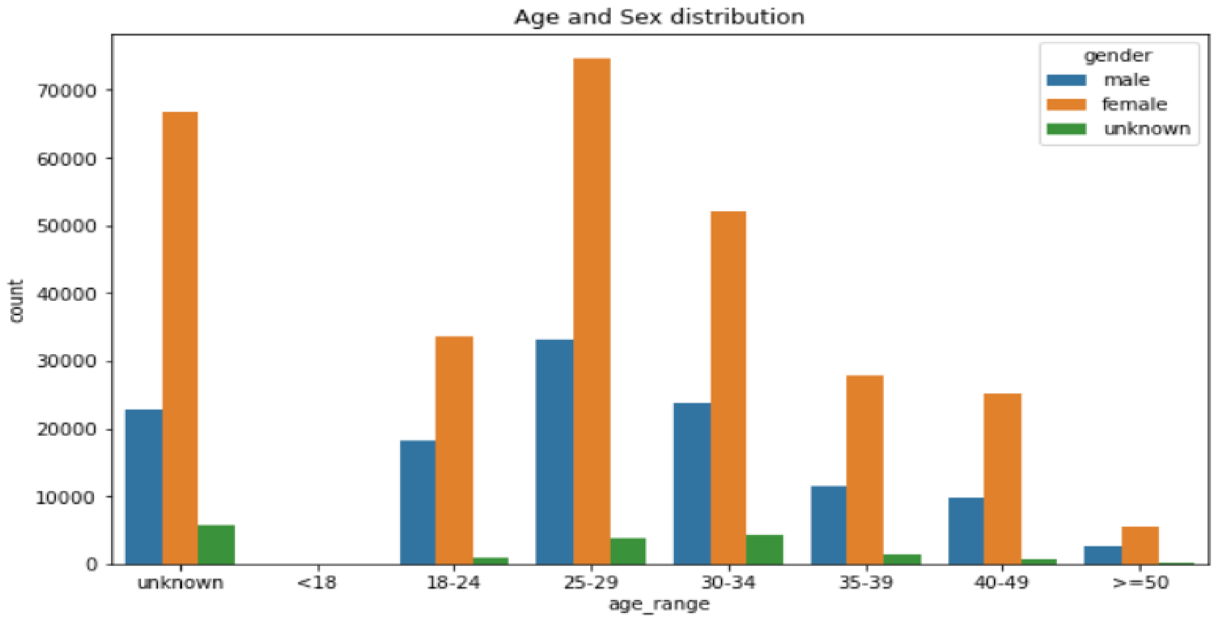
Exploratory Data Analysis and initial findings

a) Value counts of label



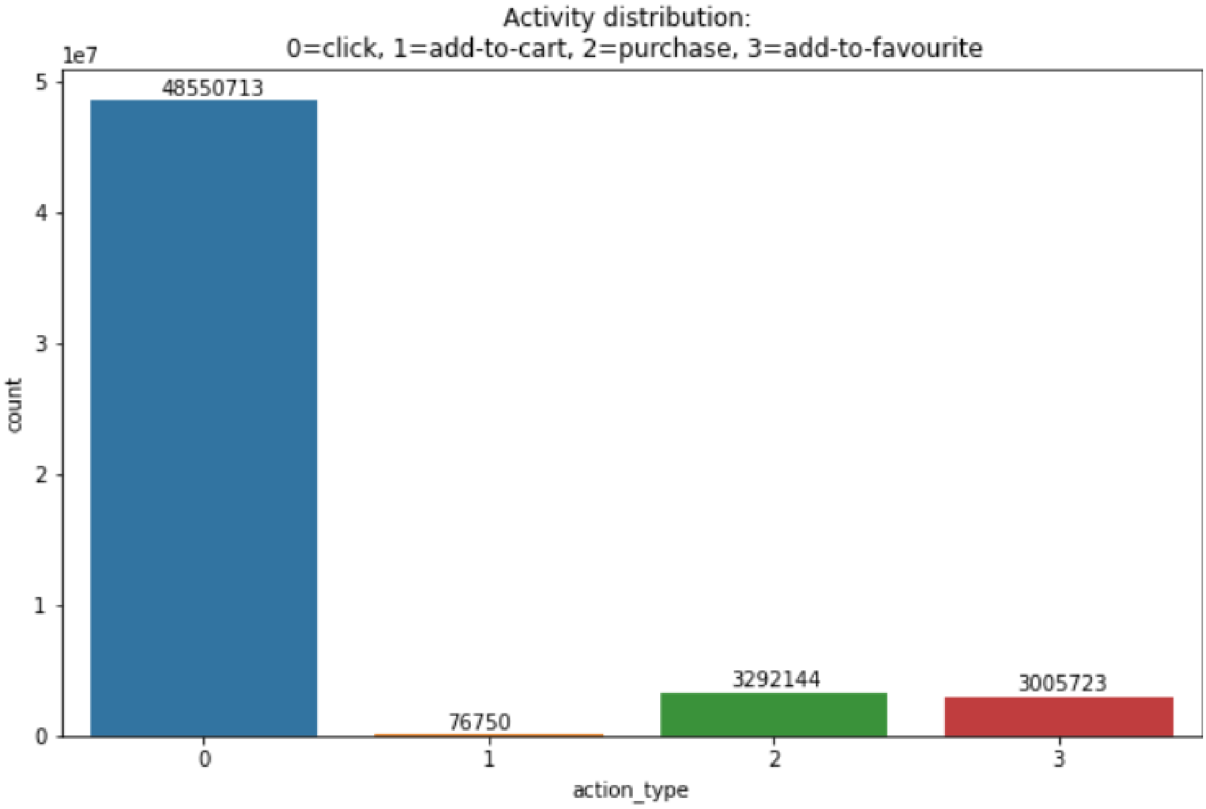
Findings: around 15:1 ratio between non-repeated customers and repeated customers, which is highly imbalanced

b) Plots of customer distribution by age and gender



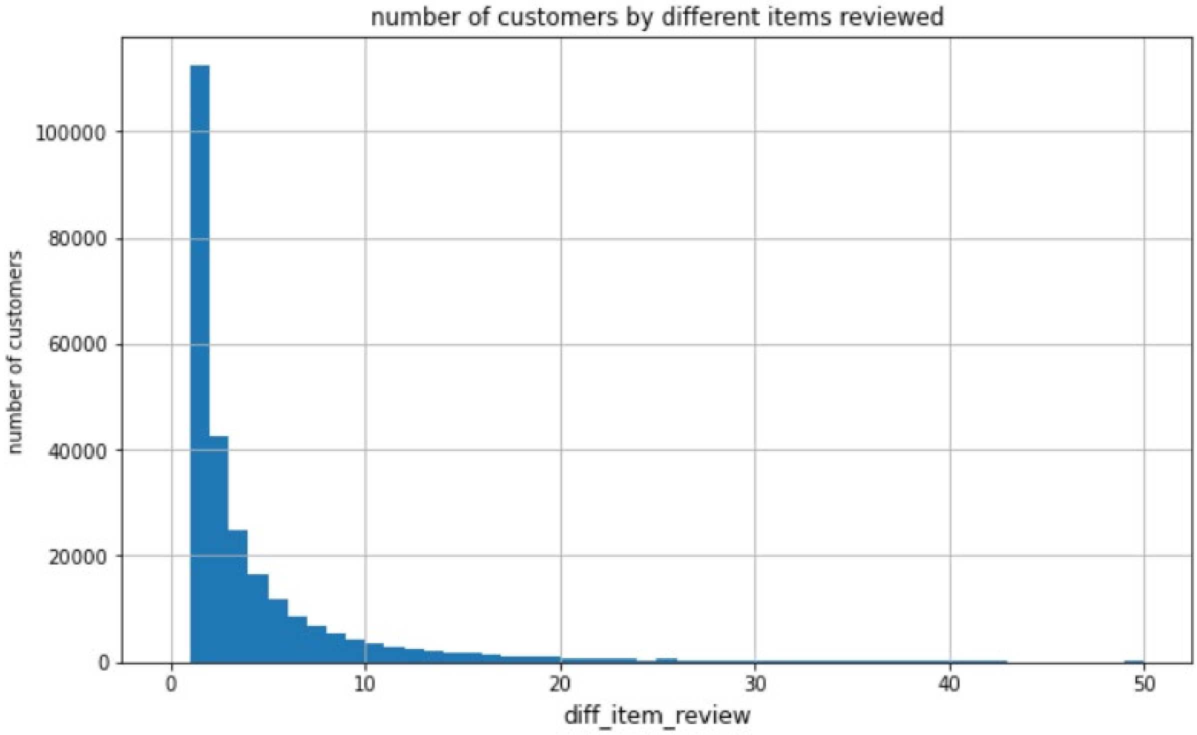
Findings: majority of customers are aged between 18 - 34, and much more female customers than male.

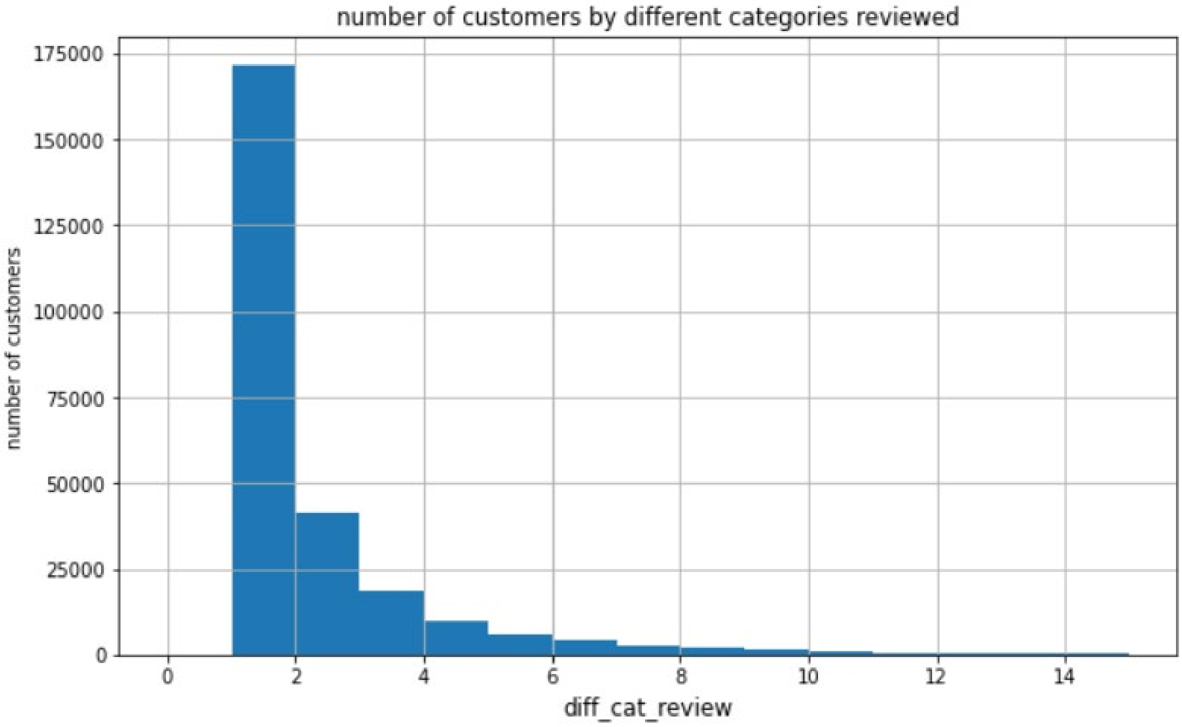
c) Plots of activity distribution



Findings: most of the time, customer only clicks on the item without any further actions.

d) Bar plot of customer counts by different items reviewed and by different categories of items reviewed



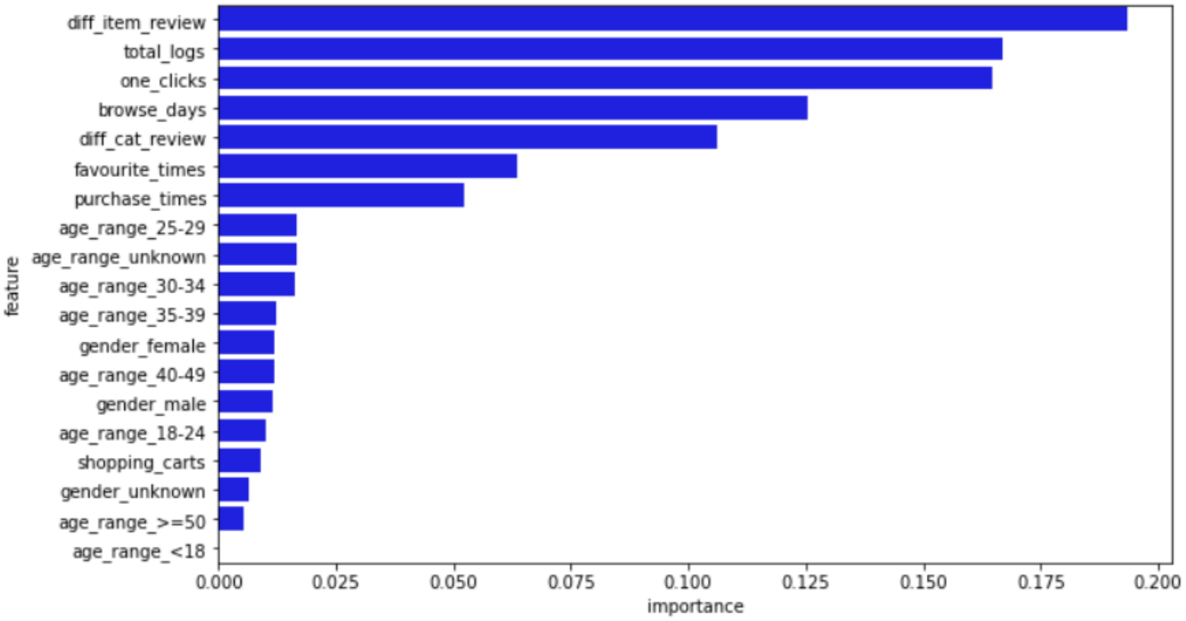


Findings: most customers reviewed less than 20 items in less than 10 different categories, indicating that majority of customers have their pre-decided target items to review or purchase; from the aspect of platform and merchant, attraction of ads and referral of traffic need to be improved

Preprocessing

First, I dropped user_id and merchant_id columns since these are only used for identification purpose not for modeling. Furthermore, I defined column 'label' as the predicting target (y) and remaining columns as predicting features (X).

Then, I fitted the dataset to a default Random Forest model and plotted the feature importance list in order to get a preliminary view on how it looks.



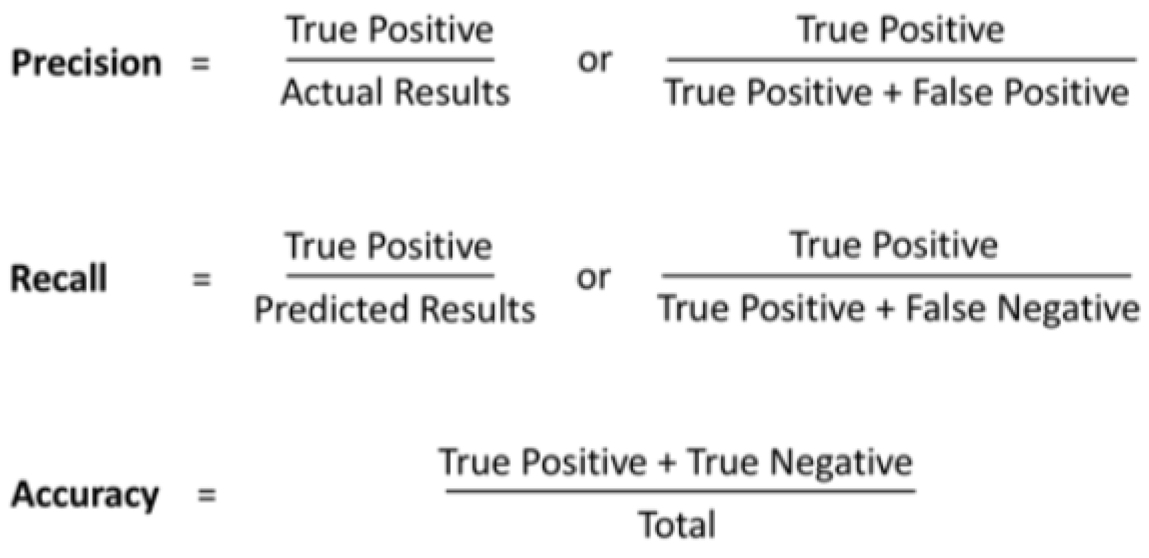
The feature importance plot indicates that the actual activity features are more important than the customer's bio features in predicting the repeated buyers, which make sense.

Due to the fact that the dataset is highly imbalanced, after splitting the dataset into training set and test set, I used up-sampling to balance the training set while keep the test set unchanged.

Modeling

I selected 4 algorithms, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Trees, for this classification capstone project.

For an imbalanced training set, the regular accuracy measure may not be the best way to evaluate the performance of a classification algorithm. For instance, when predicting the potential customers that will purchase from the merchant in the future, the recall measure seems more important than precision measure since the merchant values more about figuring out those potential customers as much as possible rather than keeping a high correctness rate.



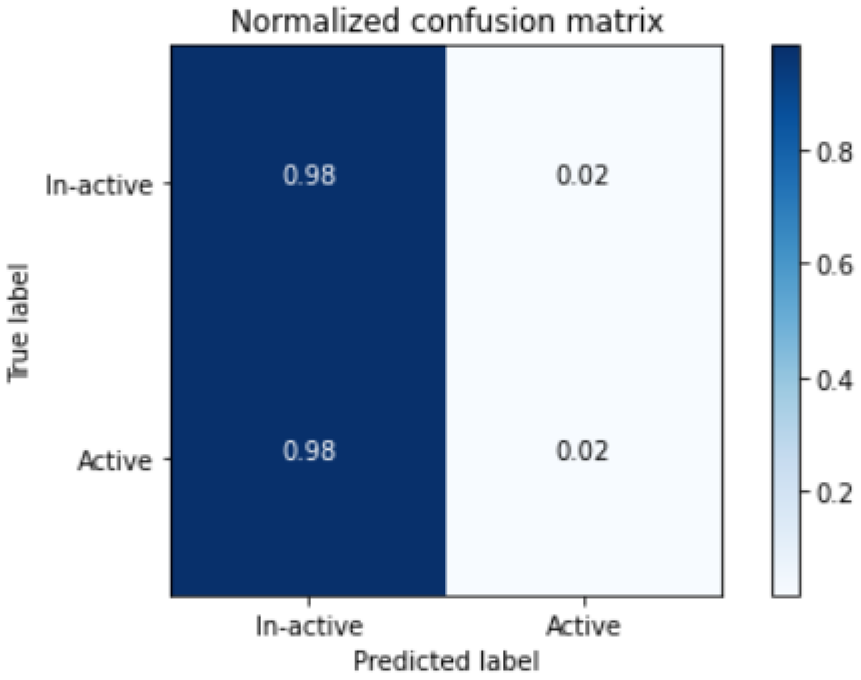
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

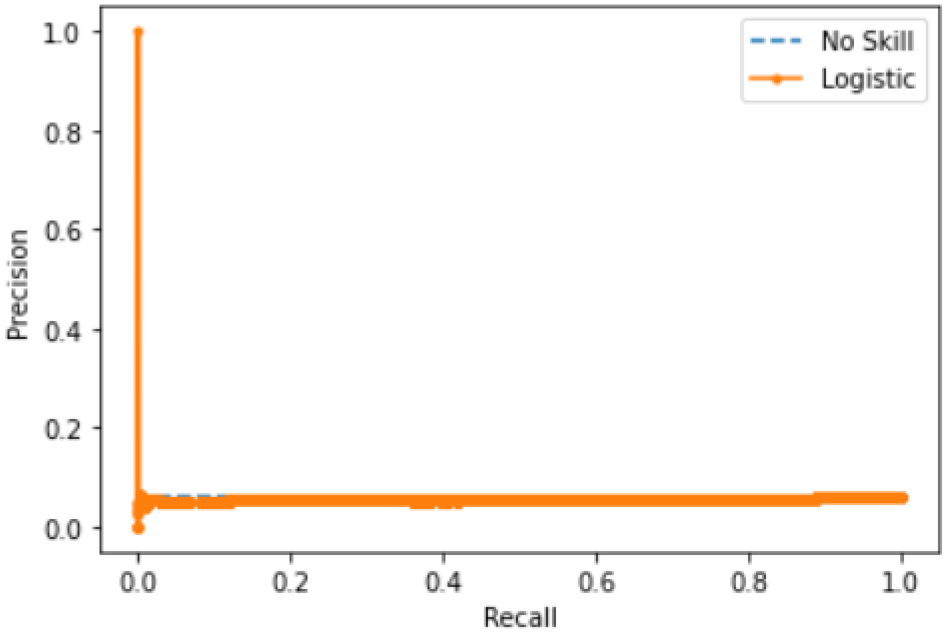
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

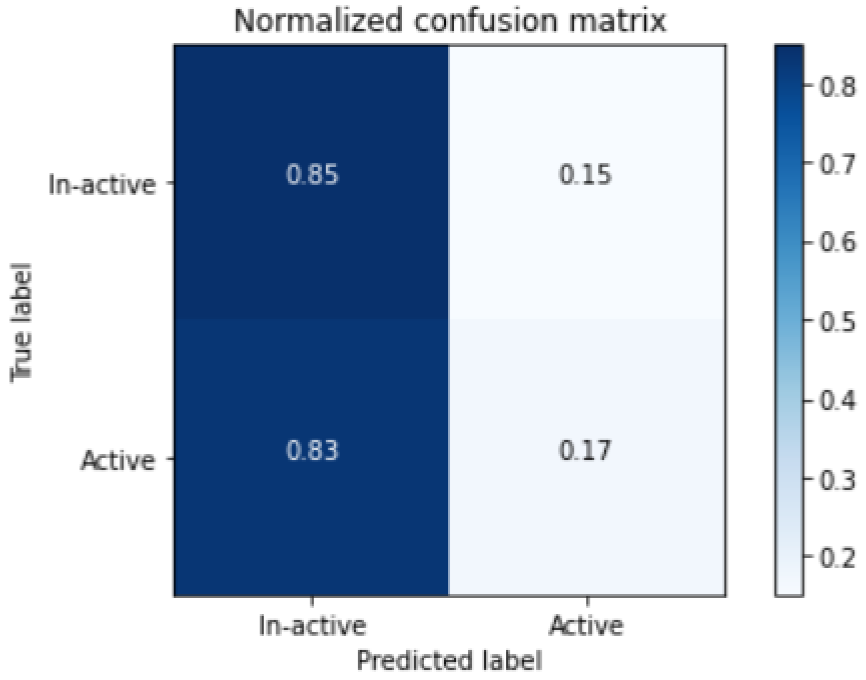
The confusion matrix and precision-recall curve plots for each algorithm are shown as below:

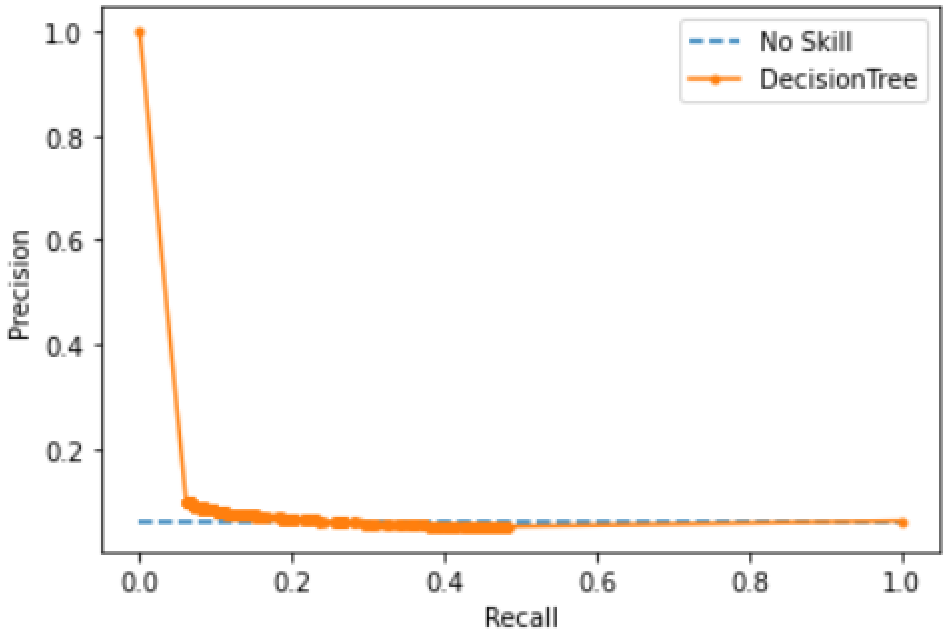
1. Logistic Regression



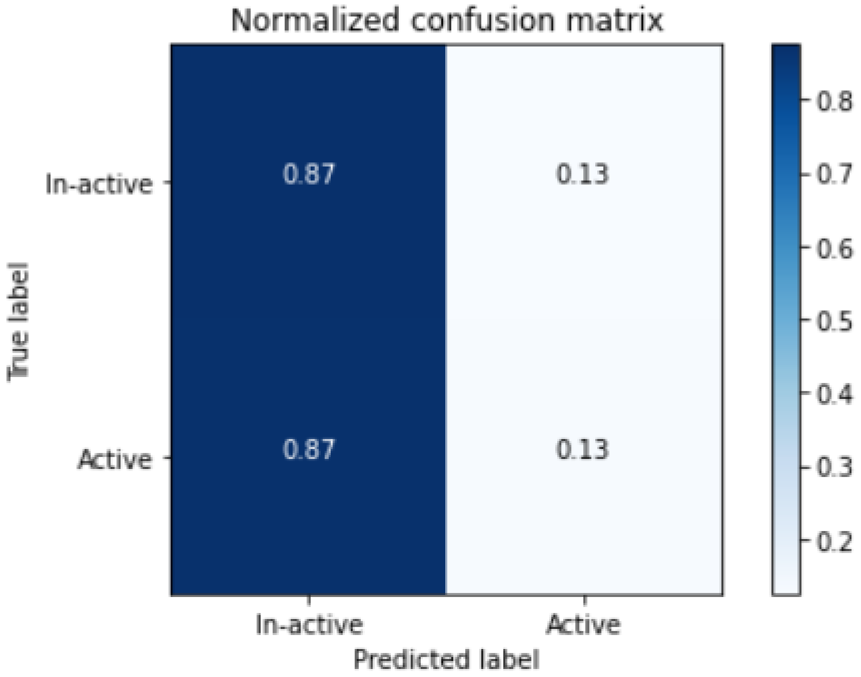


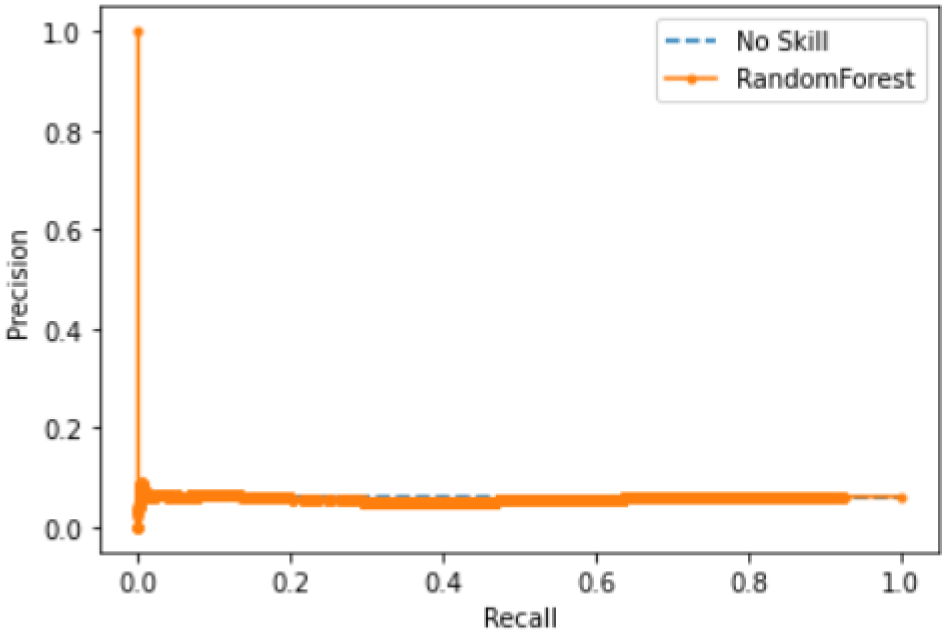
2. Decision Tree



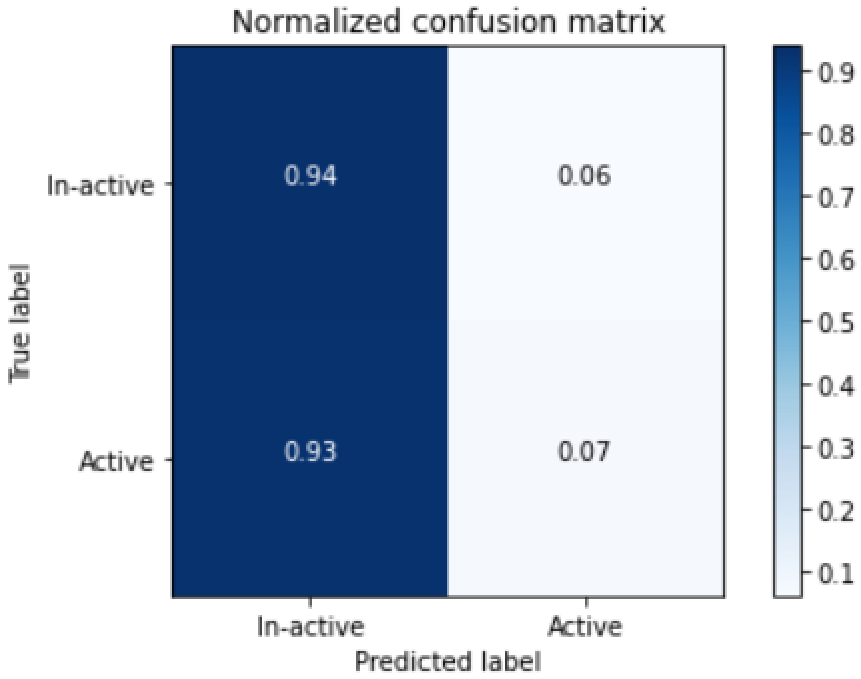


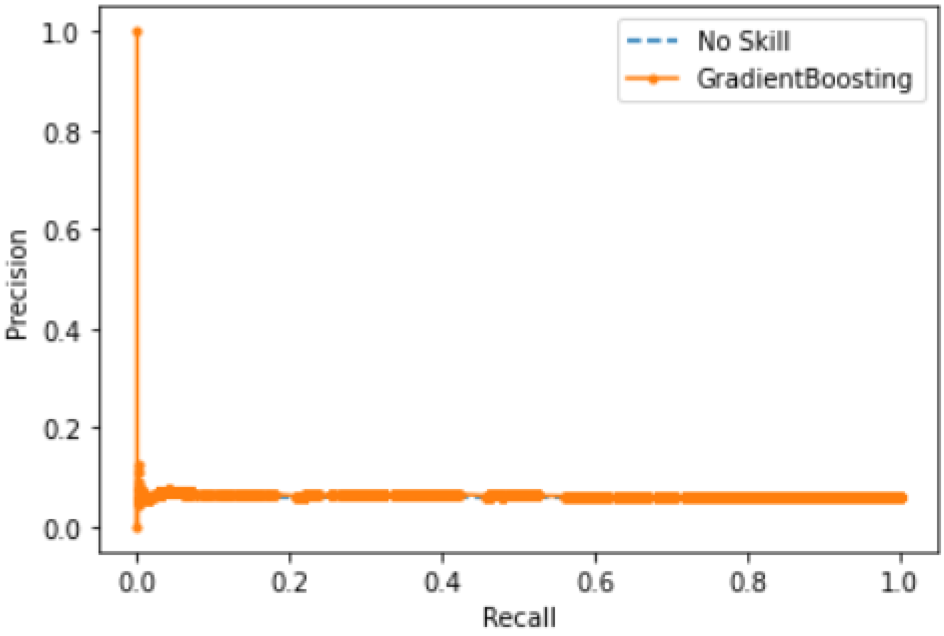
3. Random Forest





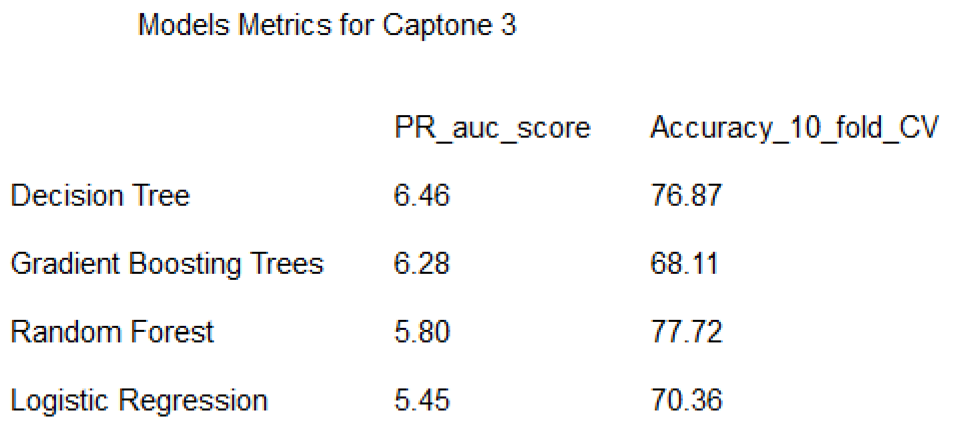
4. Gradient Boosting Trees





Metrics used for evaluation are:

1. Precision recall score (PR_auc_score)
2. Accuracy (10-fold CV)



Models Metrics for Captone 3

	PR_auc_score	Accuracy_10_fold_CV
Decision Tree	6.46	76.87
Gradient Boosting Trees	6.28	68.11
Random Forest	5.80	77.72
Logistic Regression	5.45	70.36

Takeaways

Based on the results of the metrics used, Decision Tree model resulted in the best precision recall score and the second-best accuracy, so I will choose Decision Tree for overall optimal model.

Future Research

This capstone project gave me a lot to think about feature selection and creation. For instance, I would like to expand these models with more features about the product information, such as price and discount rate. Also, trying more models, such as SVM and Neural Network, and tuning more on the model parameters would be always useful for optimizing the results.