Cosmo's Customer Purchase Behavior

DEC 520Q Section C Group 13
Team Members: Micole Wen (mw481), Sarah Lee (sel61), Zareena Hameed (zh129), Kexin Mo (km581), Pranay Ganesh koneni (pk134), Benazir Khurshid (bk190)

Contents

Business Understanding

In the modern digital age, consumers prefer to shop online from the comfort of their homes or offices rather than shopping at a physical store. The e-commerce market is growing rapidly, and consumers actively seek more engaging and highly personalized shopping experiences. Online retail businesses need to have a deep understanding of and build solid connections with the customers to thrive in the highly competitive industry. Therefore, for modern online retailers, data mining is crucial to measure and analyze consumer behavior, which can empower the business to target customers better and improve the efficiency of logistics management system, thereby increasing sales and revenue.

In this case, we aim to help a medium-size cosmetics online store better understand its customers' purchasing activity, customer value and enhance sales. To achieve the business goal, we will focus on 3 core tasks:

Core Task 1: Use classification to understand what affects purchase decision and make prediction on customer's purchase decision

This task is supervised learning. The dependent variable is purchasing activity (binary, 1=purchase, 0=not purchase), and the independent variables are price, hour, abandonment rate, average number of actions. We will use Radom Forest as the modeling method.

Core Task 2: Use clustering to group customers based on their value

This task is unsupervised learning. We will use K-means to create 4 clusters for each of the following features: recency, frequency, monetary (RFM). Then we use the clustering results to segment customers into high value, medium value and low value.

Core Task 3: Use classification to predict customers' next purchase day range

This task is supervised learning, using dependent variables "NextPurchaseDay" which is the number of days between the last purchase in the past-proxy (Period: Oct-Dec 2019) data set to the next purchase in the future-proxy (Time Period: Jan-Feb 2020) data set and "NextPurchaseDayRange", a multinomial variable created over the "NextPurchaseDay"

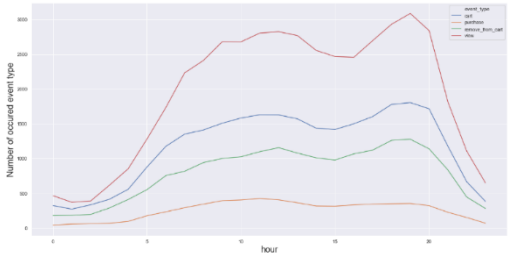
Data Understanding

Data source: [kaggle]

This dataset contains 20 million rows of users' behavior data of a large multi-category online store for five months (Oct 2019 – Feb 2020). Each row represents an event that is related to products and users. Each event has a many-to-many relation between products and users.

Column	Description
event_time	Time when an event occurred
event_type	categorizes of actions a customer made on the online store
product_id	A unique ID of a product
category_id	A unique ID of a product's category
category_code	Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories
brand	The name of a brand
price	The float price of a product
user_id	A unique permanent ID for a user
user_session	A temporary ID for each user's session. A new ID would be created when every time a user comes back to the online store from a long pause

Before data cleaning, we performed database query to draw some preliminary understanding (Appendix 1) :



- Noon and 7 pm have the highest engagement from the online store
- Total monthly sales and average spending per customer were highest in November
- Number of customers spiked at the end of Nov 2019 (Thanksgiving) and the end of Jan 2020
- A big drop was observed in customer count on New Year's Eve

Data Preparation

First, we merged the five separate files from the previous five months into one single file. There are 20692840 rows and 9 columns in this dataset. Because the dataset is so large, we'd choose a sample size

of 100000 records at random and then check for any irrelevant entries. We opted to exclude the columns "brand," "product id," "category id," "category code," and "user session" after deleting the duplicates because these variables aren't important to our anticipated model.

We also decided to remove all records with negative prices. In our investigation, we found there are five products with negative prices in the dataset. Observing the transactions of each product with negative prices, we found they do not have any price change through months; most of the event types of these transactions are direct purchases. Hence, these transactions are not complete. For instance, buying items without adding them to the cart isn't logical. Therefore, we conclude that transactions with negative prices are possible errors instead of refunds. After the data cleaning process, we checked for the dataset's basic validation of the cleaned data frame.

Creating New Columns

The average number of actions per customer and cart abandonment rate would be used in predicting our first core task. The average number of actions per customer are calculated by the average number of activities on all products made by each user id. For example, if the average number of actions is one, the customer might be just browsing in the online store. The higher the average number of actions represents higher engagement and possibility of purchasing from the user.

Users who add items to a shopping cart but leave without purchasing anything are particularly valuable to the business because it could possibly convert into purchases. The calculation of the shopping cart abandonment rate shows as following: [1 - [completed purchases/shopping carts created]] * 100 = Shopping Cart Abandonment Rate. This metric provides a specific indication of why revenue may go up or down. This indicator is valuable to the online store in making strategic decisions of converting online visitors to customers.

To prepare for achieving our second and third core tasks, we utilized the technique RFM analysis to quantitatively segment customers based on the recency, frequency, and monetary total of their transactions from Oct 2019 to Feb 2020.

Based on the individual Recency, Frequency, and Monetary clusters, an overall score is calculated. These scores are then used for segmenting customers into Low Value, Mid Value and High Value buckets. Inspired by Chen (2012), we decided to do RFM analysis by K-Means clustering. We would only be using the records with customers who made at least three purchases. The first two things we did were standardizing the data and finding the optimal number of clusters by the elbow method. Then, we would use RFM to predict future purchases from customers.

In preparing the prediction for our third model, we created the columns "DayDiffMean" which is calculated as the average number of days between each purchase. We also created "NextPurchaseDayRange" to reorganized customers' purchasing range into three categories as following: less than 20 days (tagged as active), between 20 and 40 (tagged as regular), greater than 40 days (tagged as inactive) (Appendix 2).

Modeling and Evaluation

Model 1: Random Forest

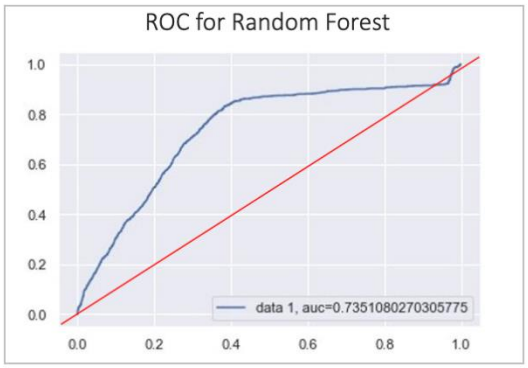
A Decision Tree is the building block of a random forest model, as it is in simple words an aggregation of the prediction/classification a decision tree would give out for a binary/categorical dependent variable. The series of Yes/No or 1/0 from multiple decision trees built using multiple combinations of the independent variables will result in the final prediction, which usually is better than or relatively more accurate than the individual decision tree.

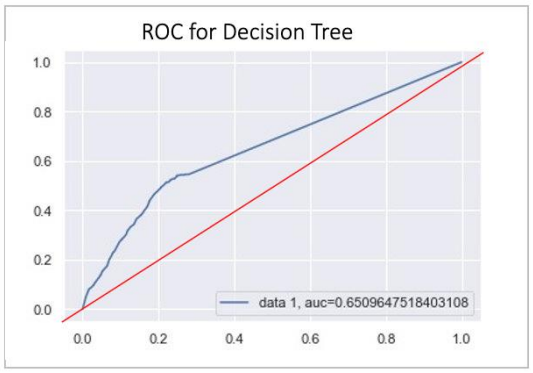
Also, as the average of the predictions only justified the word "forest", the word "Random is justified by the fact that the sampling of the training data is randomly done, so are the subset combinations of the features used for splitting at the nodes.

Random forest is better than a simple decision tree, because of the simple logic that the average of multiple individual predictions is better than relying on just one of those predictions.

First, we standardize the data. Then, we are using Random Forest here to predict the binary classification where 1: the customers made a purchase and 0: means the customer didn't make any purchase at all.

To start with, we randomly selected 80% of the data for training and set-aside the rest 20% for testing/validation once the model is built to evaluate the error*accuracy of the predictions from the model.





	RFM	Decision Tree
Accuracy	93.24 %	94.45%
Sensitivity	6.87 %	15.5%
Specifity	97.8 %	97.62%
AUC	0.73	0.65

We compare the performance of the RFM and Decision Tree. The two models have similar accuracy and specifcity, and RFM has a better AUC. RFM has a better performance overall, so we would use this model as our prediction model (Appendix 3).

Model 2: RFM + K-Means

In this model, we are trying to analyze the past purchasing behavior of the customers to predict the future purchasing pattern, which is basically predicting just "how many days the customer will take to make the next purchase in future". To. To build this model, we need to create the dependent variable in the existing data as "number of days the customer took before making the next purchase".

Creating the dependent variable:

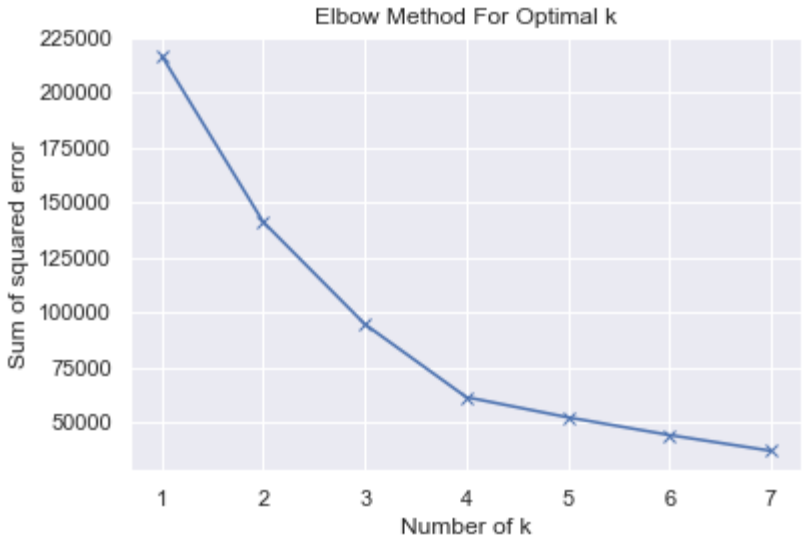
To create this dependent variable, we are dividing the original data set into 2 parts based on the time period:

- → Data Period 1. Time period selected, to be the past purchasing-behavior: Oct, Nov, & Dec 2019
- → Data Period 2 Time period selected, to be used as future purchasing-pattern: Jan, & Feb 2020

Then, make the last purchase for every customer in Data Period 1 ((Max of the purchase date for every customer in Data Period 1)) check how many days the customer took before making the first purchase in Data Period 2 (max of purchase date for every customer in Data Period 2).

Modelling steps:

- → We performed RFM (Recency, Frequency, & Monetary) and Clustering using K-Means on the variables (features)(Appendix 4).
- → The optimal number of clusters when using K-Means has been selected using the Elbow Method, which was basically calculating the distortion score on Y-Axis (sum of the squares of the distance of each point in a cluster from the centroid of their respective cluster) for multiple K values and plotting it against the K-Values on X-Axis. The optimal number of clusters to be used when performing K-means is the point (# of clusters value on X-axis), after which the distortion (not inertia in this case) starts to decrease in more of a linear fashion or flattening, relatively.
- → Then we will create 3 K-means clustering models for each of R,F, & M to segment the customers into 4 clusters, as 4 has been selected as the optimal number of clusters to be used based on the Elbow method. We then created an overall score for each customer by aggregating (sum) their RFM cluster number as absolute number to score the customers on a scale where higher the score - higher the value.



→ Then based on the score created, the data distribution based on this score variable is analyzed to identify whether a customer is of high/med/low value

Model 3: Regression & Classification

This task is supervised learning as we have already created a dependent variable called "NextPurchaseDay" which is the number of days between last purchase in the past-proxy (Time Period: Oct-Dec 2019) data set to the next purchase in the future-proxy (Time Period: Jan-Feb 2020) data set. We ran a regression model with independent variables are Recency, Frequency, Revenue, DayDiffMean and Customer Segments (Low, Medium, & High) and "NextPurchaseDay" as dependednt variable, building the model to predict the number of days a customer would take to make the next puchase based on the past-behavior.

However, the results of the linear regression were discouraging. Hence, after that, we decided to attempt with the classification models. For that, we created a new categorical variable "NextPurchaseDayRange" based on the existing continuous variable "NextPurchaseDay", tagging the customers as inactive, active, & regular based on the data-distribution analysis(Appendix 5, 6).

Classification Modeling(Appendix 7):

We first started by running several Classification models and selected to go ahead with attempting,

based on F1 Score, Accuracy, Training Time, Testing Time:

	Accuracy	Cross Validation Accuracy
SVM	60.46 %	54.76 %
RF	60.78 %	52.89 %
XGB	55.56 %	50.87 %
AUC	0.73	0.65

Deep Learning

Classification Report:

	precision	recall	f1-score	support
0	0.58	0.93	0.71	194
1	0.22	0.10	0.14	96
2	0.00	0.00	0.00	70
accuracy			0.53	360
macro avg	0.26	0.35	0.28	360
weighted avg	0.37	0.53	0.42	360

So, based on the above model results and evaluation metrics, we decided to choose RF as the best predictive model for the business problem.

Deployment

Core Task 1:

The result of core task 1 indicates that price and shopping hours are important to our customers. Therefore, we recommend the business be sensitive to its pricing strategy and advertising hours. First, business can use the model to track the prediction of customer purchase decisions. For those observations shown as "not purchase", it can dig in to see if there are patterns across brands or product categories. If a certain brand or category pops out, the business should check its pricing strategy to see if any optimization can be applied (We are not given brand and category info in this case, but the business can apply this in the future). Second, if the business wants to buy advertising for the brand or category that

detected in step 1 to boost up sales, it should deliver the advertisement around peak user active hours, which are noon and 19 pm.

Core Task 2:

The result of core task 2 helps the business to segment customers based on values. If the business plan to initiate a loyalty program for customer retention, this segmentation can be used to determine offers differentiation across the customers to achieve higher cost-efficiency on customer retention. For the high-value customers, the business should treat them with the highest priority because the business should aim at retaining them in the same segment so that they continue to offer high value to the business. For medium-value customers, the business should show them attractive incentives of moving to higher loyalty levels, such as more privilege, to motivate them to convert to high-value consumers. For the low-value customers, the business can evaluate their future worth to determine if they are worthy of retention and business's investment. The low value customers are worth analyzing more since its cheaper to retain these customers than to acquire new customers. (We are not given a sufficient period of data to predict customers' future value in this case).

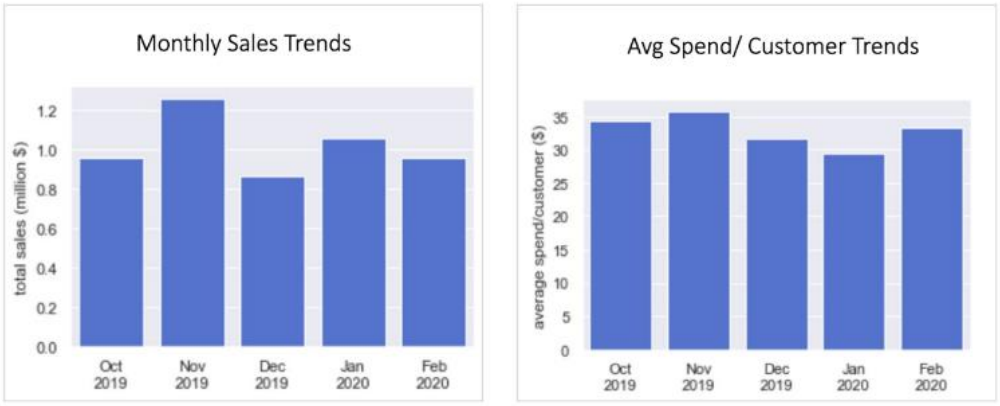
Core Task 3:

With the next purchase day range prediction, the business can run retention campaign targeted at those inactive customers. For example, sending them personalized reach-out including call to action, discount incentive, or creative trigger. In addition, the classification can be used to help inventory management. After generating the next purchase day range, the business can deep dive into the three different ranges and look for patterns of the observations, such as geographic, product origins, product categories, brands, etc. (we are not given these variables in this case). Then the business can develop inventory management strategies based on the demand prediction. For example, if the business analyses the observations which are classified as "will have purchase less than 20 days" and finds a high percentage of certain Japanese Brand, it should be aware of the inventory demand of this brand and keep enough inventory to provide customer best shopping experience.

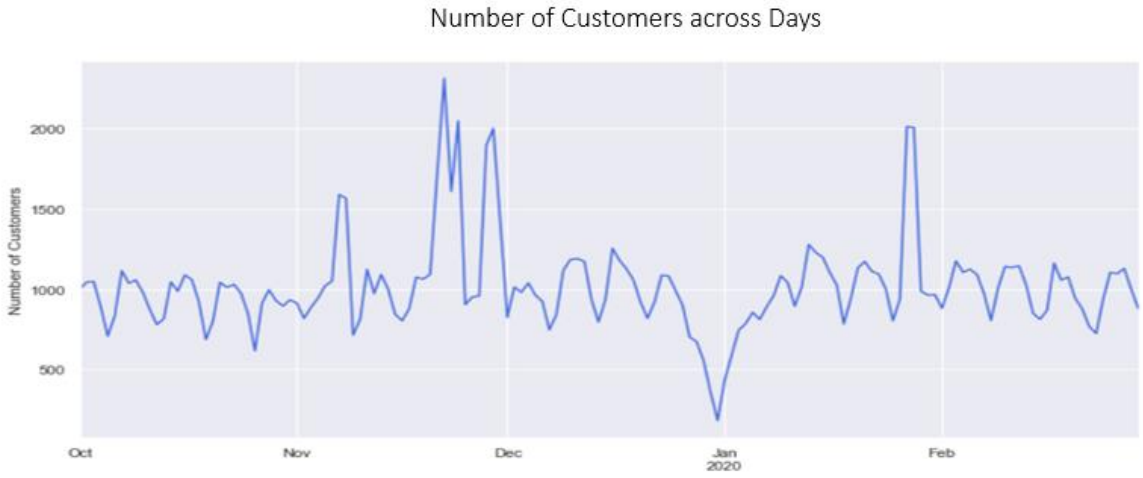
Appendix

1. Plots for data exploration

1) The Histogram of Monthly Total Sales per Month/Avg $ Spent Per Month



2) Number of Customers across Days



2. Descriptive Statistics for Numerical Variables

1) Variables for Model1

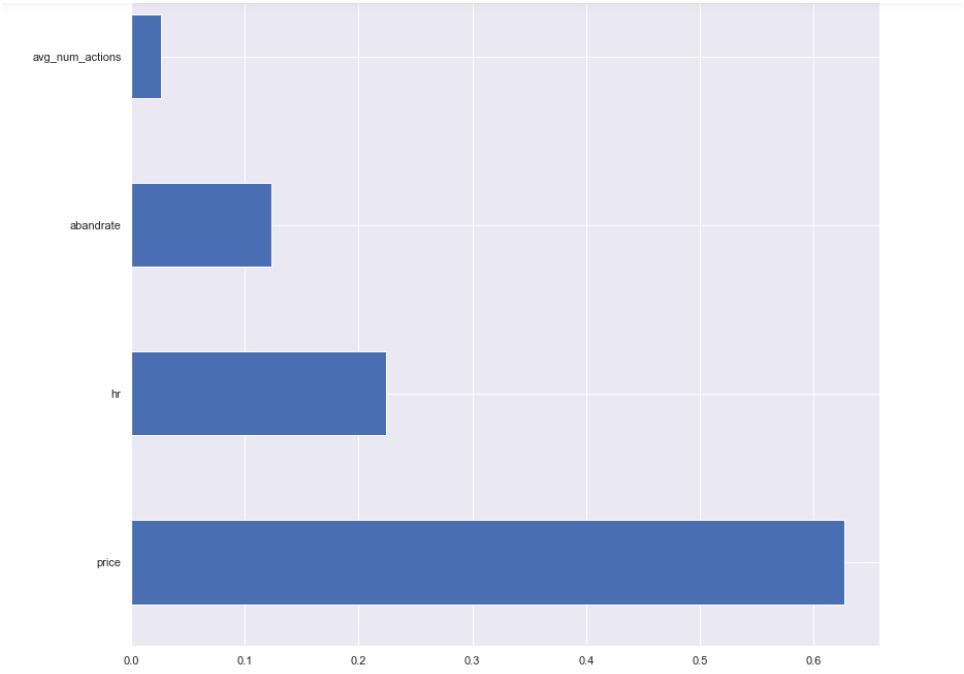
	price	avg_num_actions	abandrate	purchase
min	0.000	1.000000	-4.000000	0.0000
max	327.780	3.500000	5.000000	1.0000
mean	8.447	1.011612	0.978149	0.0499
50%	4.020	1.000000	1.000000	0.0000

2) Variables for Model2

	NextPurchaseDay	Recency	Frequency	Revenue	DayDiffMean	NextPurchaseDayRange
min	2.00000	0.000000	4.000000	9.670000	1.333333	0.000000
max	999.00000	79.000000	398.000000	2497.860000	30.333333	2.000000
mean	335.90995	18.576987	52.011673	245.807487	14.748442	0.647026
50%	42.00000	15.000000	44.000000	199.030000	14.600000	0.000000

3. Calculating the importance of the features picked up by the model:

-> Higher weightage is attached to the feature (variable/node), which sees higher % of the total sample size passing through.



4. Clustering Customers based on RFM value using K-means

RecencyCluster	count	mean	std	min	25%	50%	75%	max
0	14598.0	79.998219	6.684710	69.0	74.0	80.0	86.0	91.0
1	15680.0	56.419005	6.478197	46.0	51.0	56.0	62.0	68.0
2	21778.0	34.679952	5.714011	24.0	31.0	35.0	39.0	45.0
3	20169.0	12.577520	6.314917	0.0	7.0	13.0	18.0	23.0

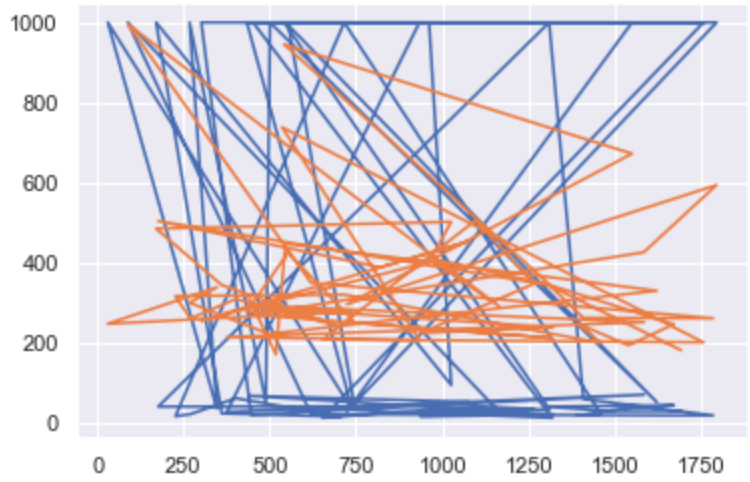
FrequencyCluster	count	mean	std	min	25%	50%	75%	max
0	53900.0	4.945269	3.158669	1.0	2.0	4.0	7.0	12.0
1	14730.0	20.209437	6.225539	13.0	15.0	19.0	24.0	36.0
2	3247.0	52.652294	13.622094	37.0	41.0	49.0	60.0	91.0
3	348.0	130.362069	46.860833	92.0	100.0	116.5	139.0	438.0

RevenueCluster	count	mean	std	min	25%	50%	75%	max
0	56351.0	27.649883	15.890115	0.13	14.0100	24.000	40.200	65.35
1	12761.0	102.899124	30.613030	65.36	77.7800	94.350	122.660	183.26
2	2814.0	263.239204	70.762454	183.27	207.6925	240.125	301.855	478.54
3	299.0	693.954615	292.666248	480.62	531.7400	596.240	747.765	2715.87

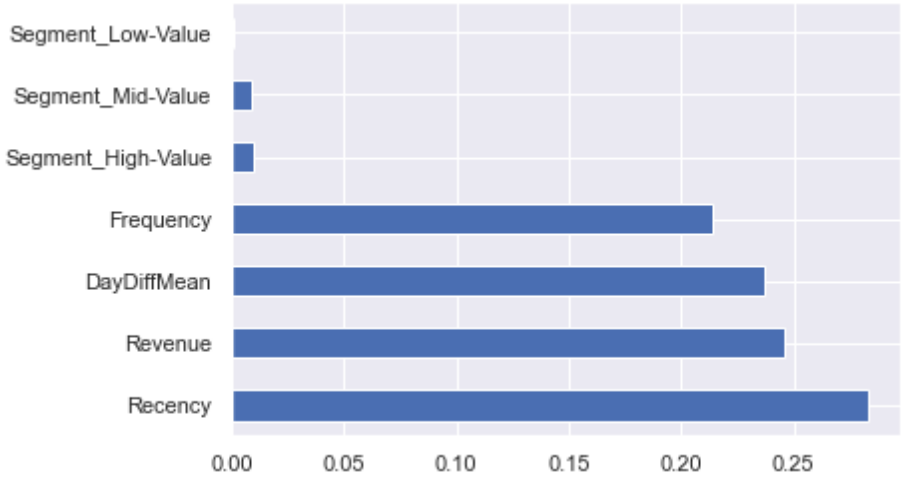
5. Coefficients for the variable used for regression(DV:NextPurchaseDate)

	Coefficient
Recency	8.179411
Frequency	0.320606
Revenue	-0.174748
DayDiffMean	-2.008662
Segment_High-Value	-56.869544
Segment_Low-Value	99.369967
Segment_Mid-Value	-42.500424

6. Performance for linear regression: Accuracy Visualization (red: Actual, Green: Predicted)



7. Feature Importance based on Random Forest Model:



Contribution of Team Members –

Coding – Kexin (Data Cleaning), Sarah (Data Cleaning), Micole (EDA), Zareena (EDA), Pranay (Modelling), Benazir (Modelling)

Presentation Slides – Zareena, Benazir, Kexin

Report – Kexin, Pranay, Micole, Sarah

Presentation- Pranay, Benz