IST 718
Big Data Analytics

Final Project Report:
Black Friday

Team Members:
Inga Fleishacker
Stephen Clark Washington III
Arielle Dortch
Lauren Foltz

Summary of Findings:

- Product category is correlated to purchase amount.
- More customers are male, single, and age 26-35.
- More customers are from city category C, but more money is spent from city category B.
- Decision Tree Regressor was the best model for predicting product category, which could act as a proxy for purchase amount.
- Product category 1, 5, 8 were the most profitable.

Recommendations:

- Focus marketing efforts on product category 1, followed by category 5 and 8.
- Look into why customers from city category C aren't spending more.
- Utilize Decision Tree and Random Forest models for future data if predicting product category.

Specification

Problem:

Black Friday is known as one of the busiest shopping days due to discounts and consumers preparing for the holidays. This is normally the time where companies see their greatest sales. Consequently, it is imperative that they market the right products to the most profitable consumers. Additionally, knowing which products or which product categories bring in the most revenue will help retailers know which items to advertise more heavily.

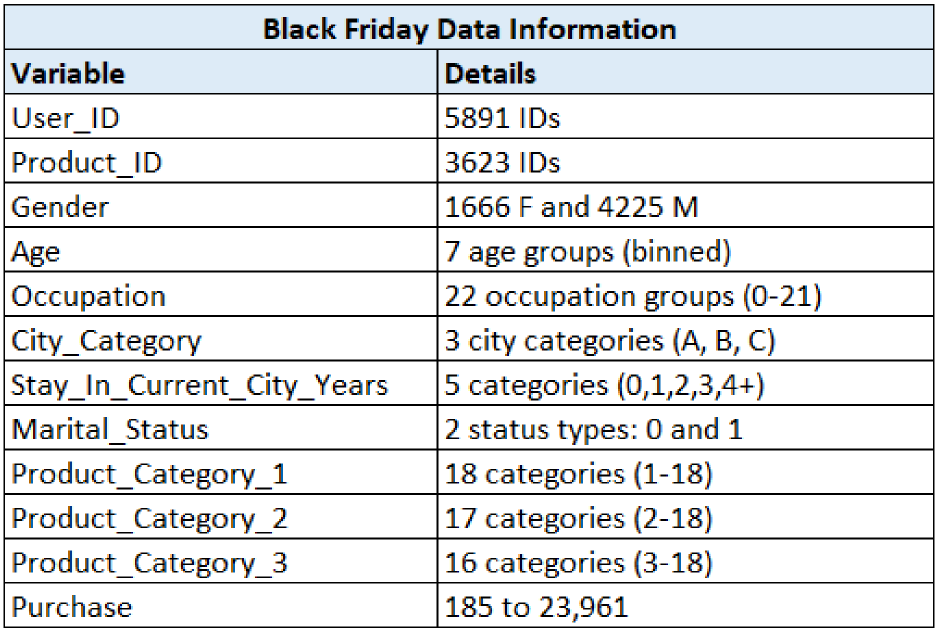
Hypothesis:

Sales are significantly correlated to specific attributes of consumers, as well as specific product types.

Data:

We chose to use Black Friday data due to it being a manageable subset of consumers' spending habits. It consists of over 500K purchases made over a one-month period, by 5,891 customers. It includes demographic information about customers, as well as information about products and spending. For confidentiality purposes, the data set masked the names of products sold, the categories under which they fall, the city category of customers and the occupation of customers. This prevented our group from making any premature assumptions. We could only infer what types of products were being sold.

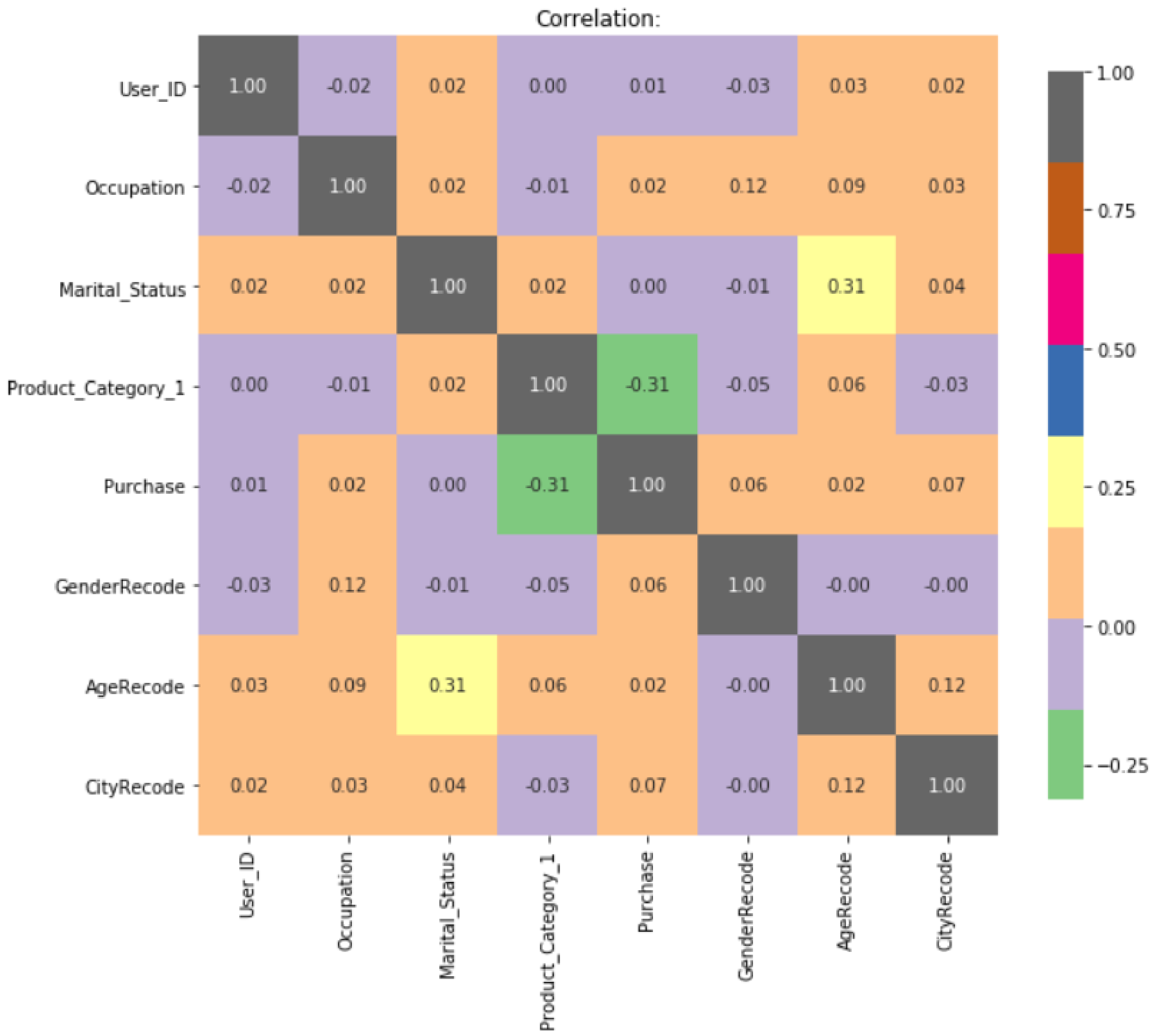
The data was fairly clean, but there were many missing values for produce category 2 (31%) and product category 3 (69%). These were categorical variables, so rather than inserting the mean values where values were missing, we decided to drop these columns.



Black Friday Data Information	
Variable	**Details**
User_ID	5891 IDs
Product_ID	3623 IDs
Gender	1666 F and 4225 M
Age	7 age groups (binned)
Occupation	22 occupation groups (0-21)
City_Category	3 city categories (A, B, C)
Stay_In_Current_City_Years	5 categories (0,1,2,3,4+)
Marital_Status	2 status types: 0 and 1
Product_Category_1	18 categories (1-18)
Product_Category_2	17 categories (2-18)
Product_Category_3	16 categories (3-18)
Purchase	185 to 23,961

Observation

A correlation matrix was generated to try to determine if any of our variables were correlated. It appears that there's a correlation between Marital Status and Age, as well as a correlation between Purchase and Product Category.



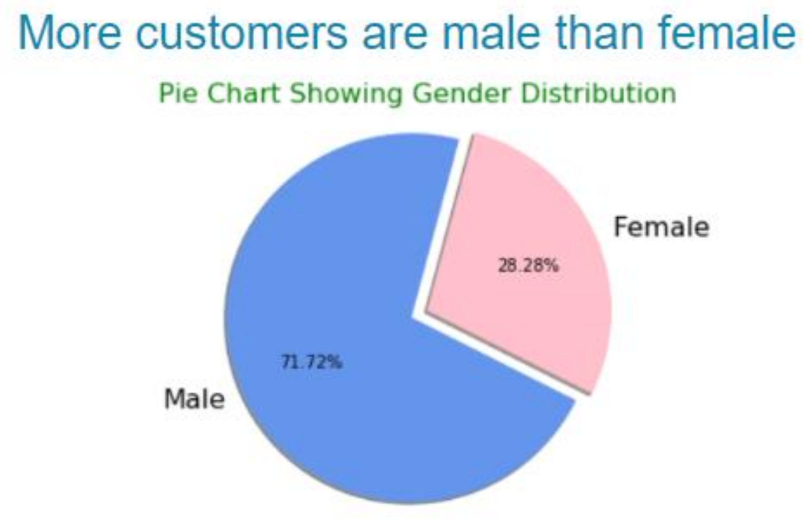
Correlation:

	User_ID	Occupation	Marital_Status	Product_Category_1	Purchase	GenderRecode	AgeRecode	CityRecode
User_ID	1.00	-0.02	0.02	0.00	0.01	-0.03	0.03	0.02
Occupation	-0.02	1.00	0.02	-0.01	0.02	0.12	0.09	0.03
Marital_Status	0.02	0.02	1.00	0.02	0.00	-0.01	0.31	0.04
Product_Category_1	0.00	-0.01	0.02	1.00	-0.31	-0.05	0.06	-0.03
Purchase	0.01	0.02	0.00	-0.31	1.00	0.06	0.02	0.07
GenderRecode	-0.03	0.12	-0.01	-0.05	0.06	1.00	-0.00	-0.00
AgeRecode	0.03	0.09	0.31	0.06	0.02	-0.00	1.00	0.12
CityRecode	0.02	0.03	0.04	-0.03	0.07	-0.00	0.12	1.00

The data consisted of multiple records per customer, so we aggregated the data by customer ID in order to get statistics about our customers. Purchase information was analyzed without aggregation, so that product category and individual purchase amount could be analyzed.

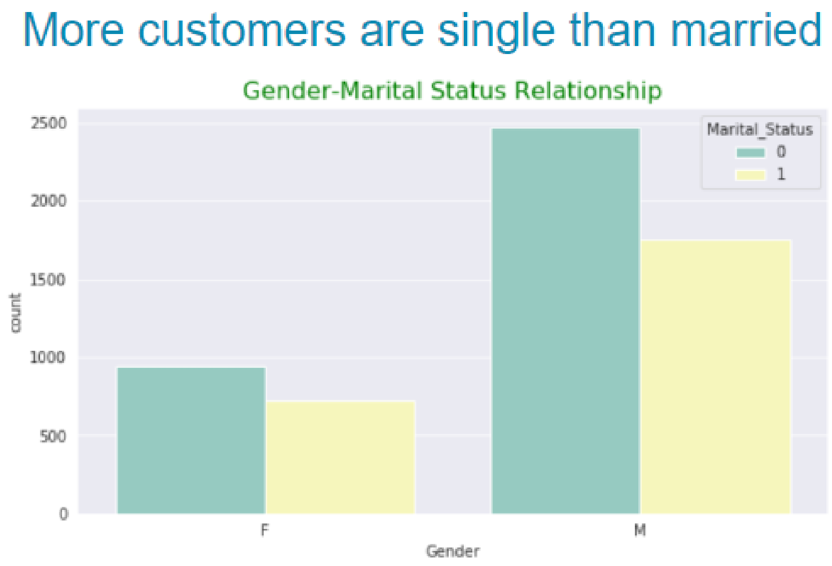
Observations on Customers:

- More customers are Male (72%) than Female (28%).
- More customers are Single (58%) than Married (42%).
- Most customers are Age 26-35 (35%).
- The predominant Occupations are 0,4,7



More customers are male than female

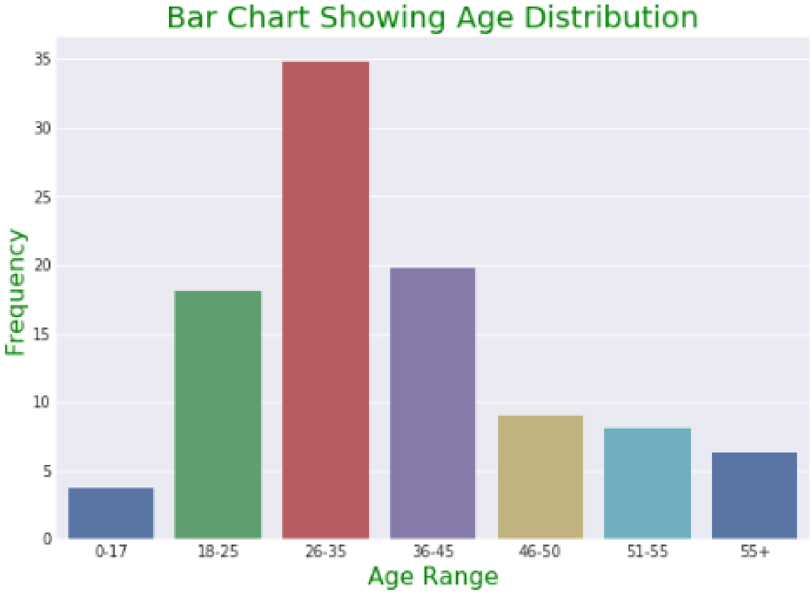
Pie Chart Showing Gender Distribution



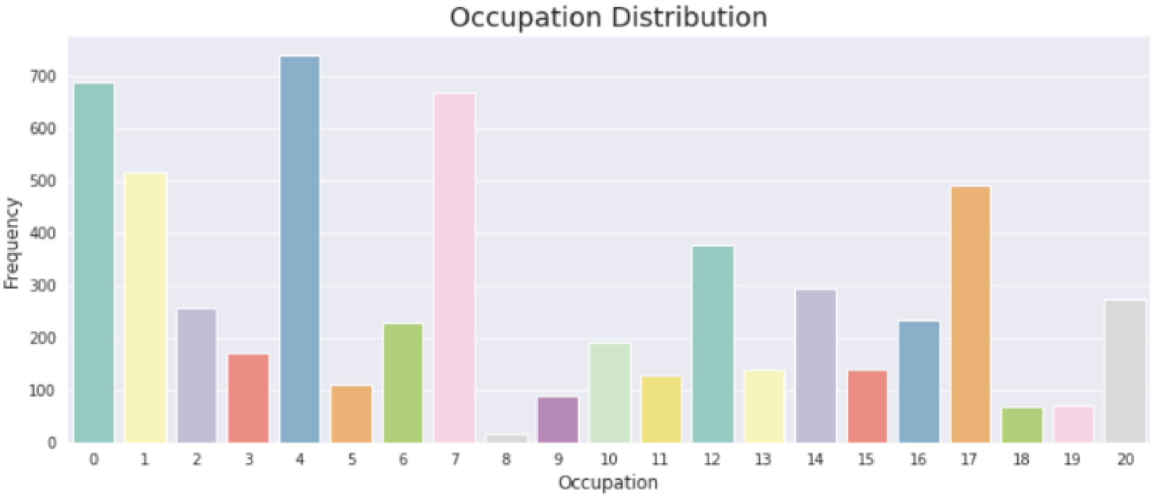
More customers are single than married

Gender-Marital Status Relationship

The majority of the customers are 26-35 years old:



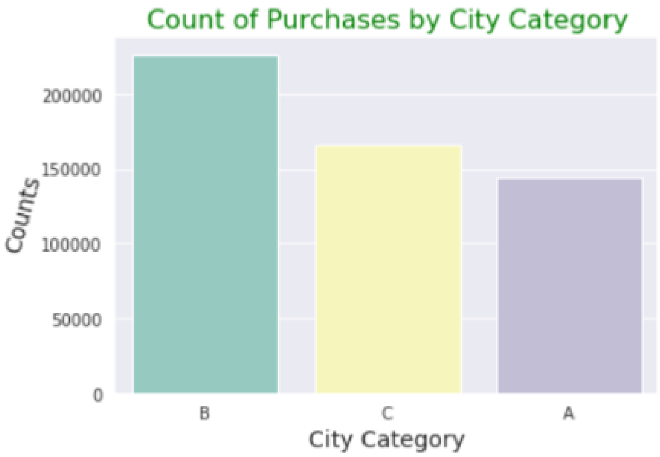
0, 4, and 7 are the most predominant occupations



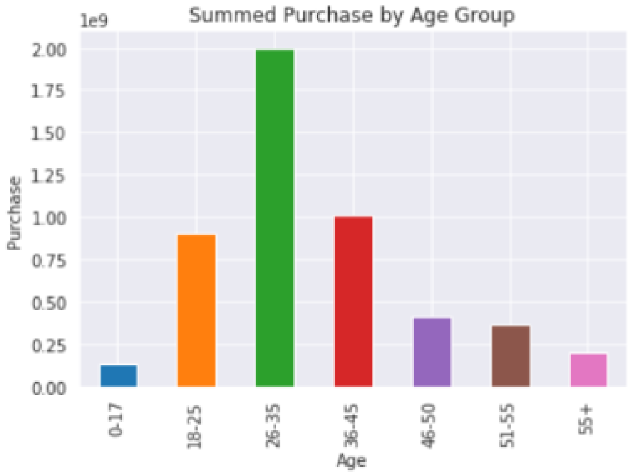
<u>Observations on Purchasing:</u>

- More purchases are from customers from City Category B
- Summed purchases are highest for Ages 26-35
- Men spent more than women in all age groups
- More money is spent by Single customers (59%) than Married (41%)

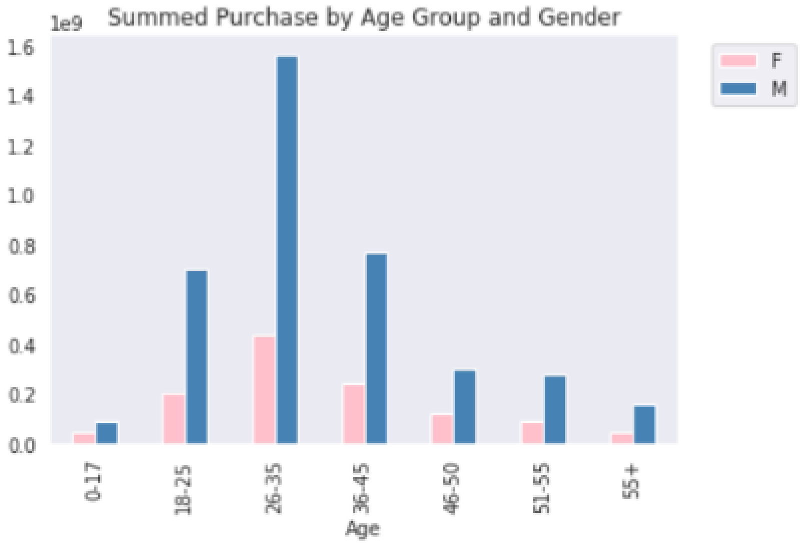
More purchases are from City Category B



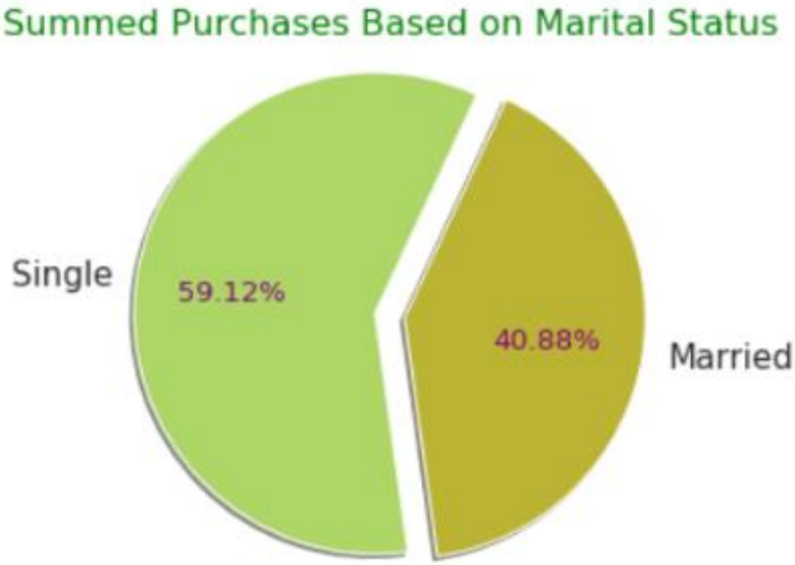
Summed purchases are highest for ages 26-35



Men spend more than women in all age groups

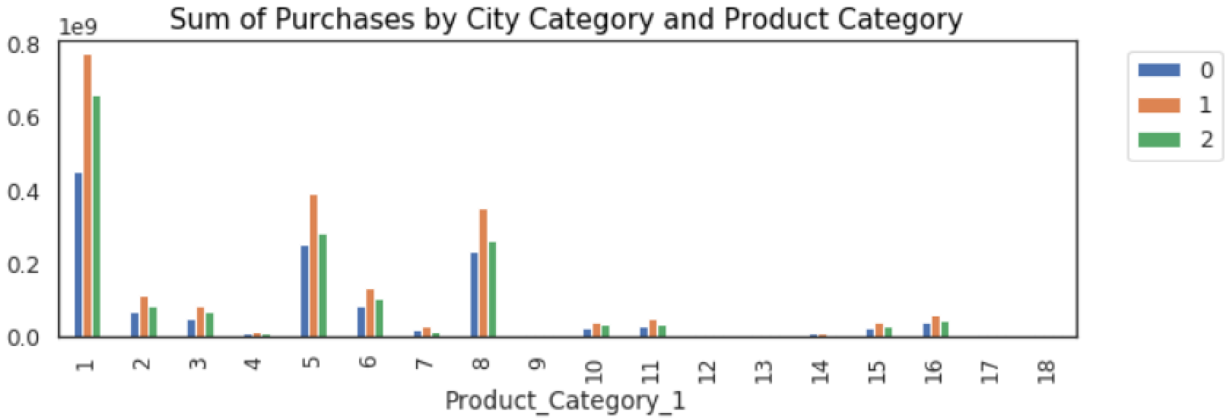


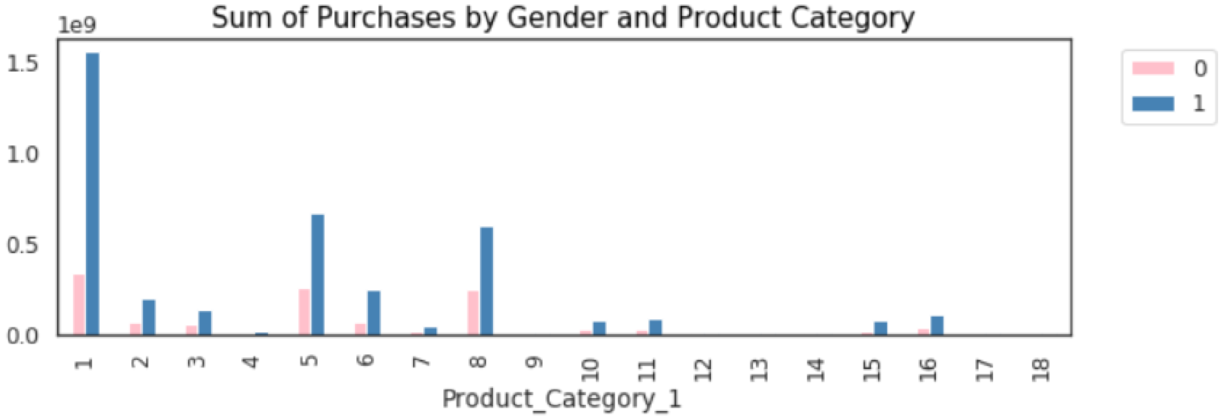
More money was spent by single customers

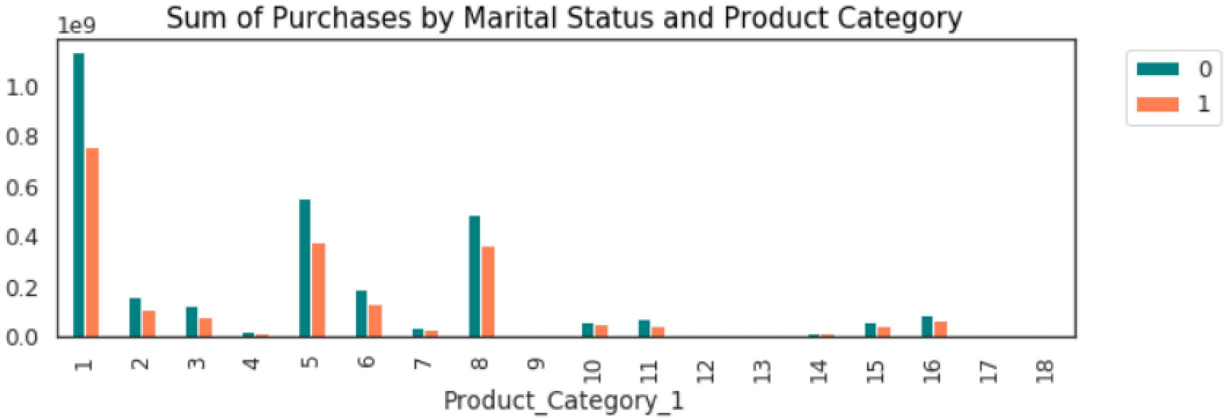


Observations on Product Categories:

The following observations compared Product Category 1 to City Category, Gender, and Marital Status. All three had similar trends. The most profitable product categories in all three charts were 1,5,8. This will help when determining which products to market for the most profit.





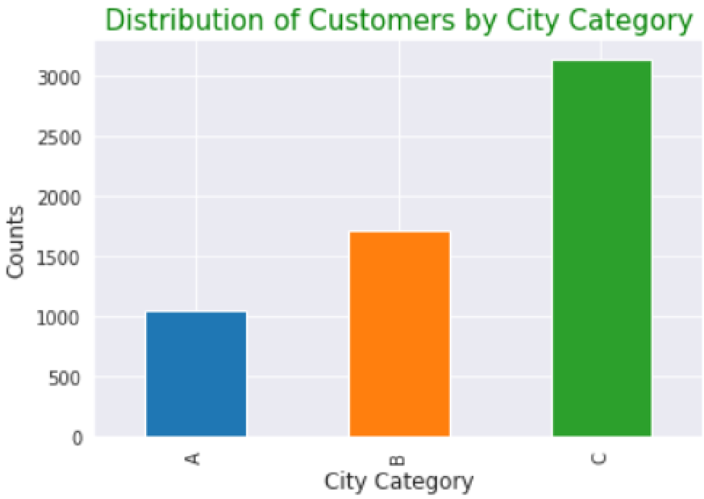


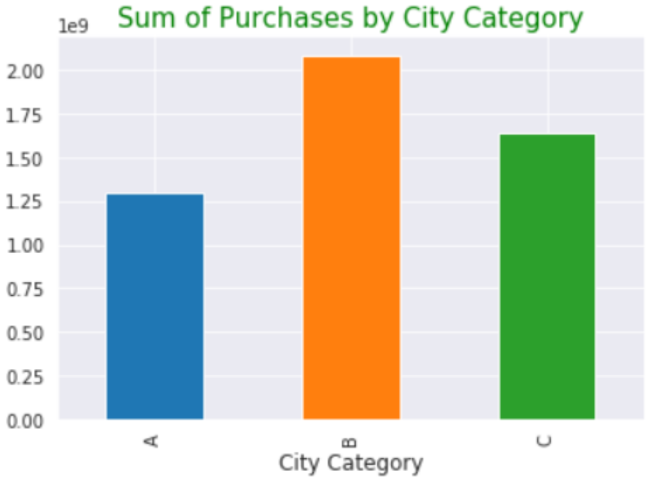
Analyses

Linear Regression

-Background: Linear regression works well for continuous data and is not computationally expensive. This was run to try to predict purchase amount, which is continuous. The goal was to determine which variables were significant.

-Findings: Linear regression was run predicting purchase amount as the target variable, and an r-squared of 62% was achieved. Significant variables were product category and city category. This gave weight to our thought that product category would be a good proxy for purchase amount as the target variable. Interestingly, even though more customers are from city C (53%), more money was spent by customers from city B (41%).

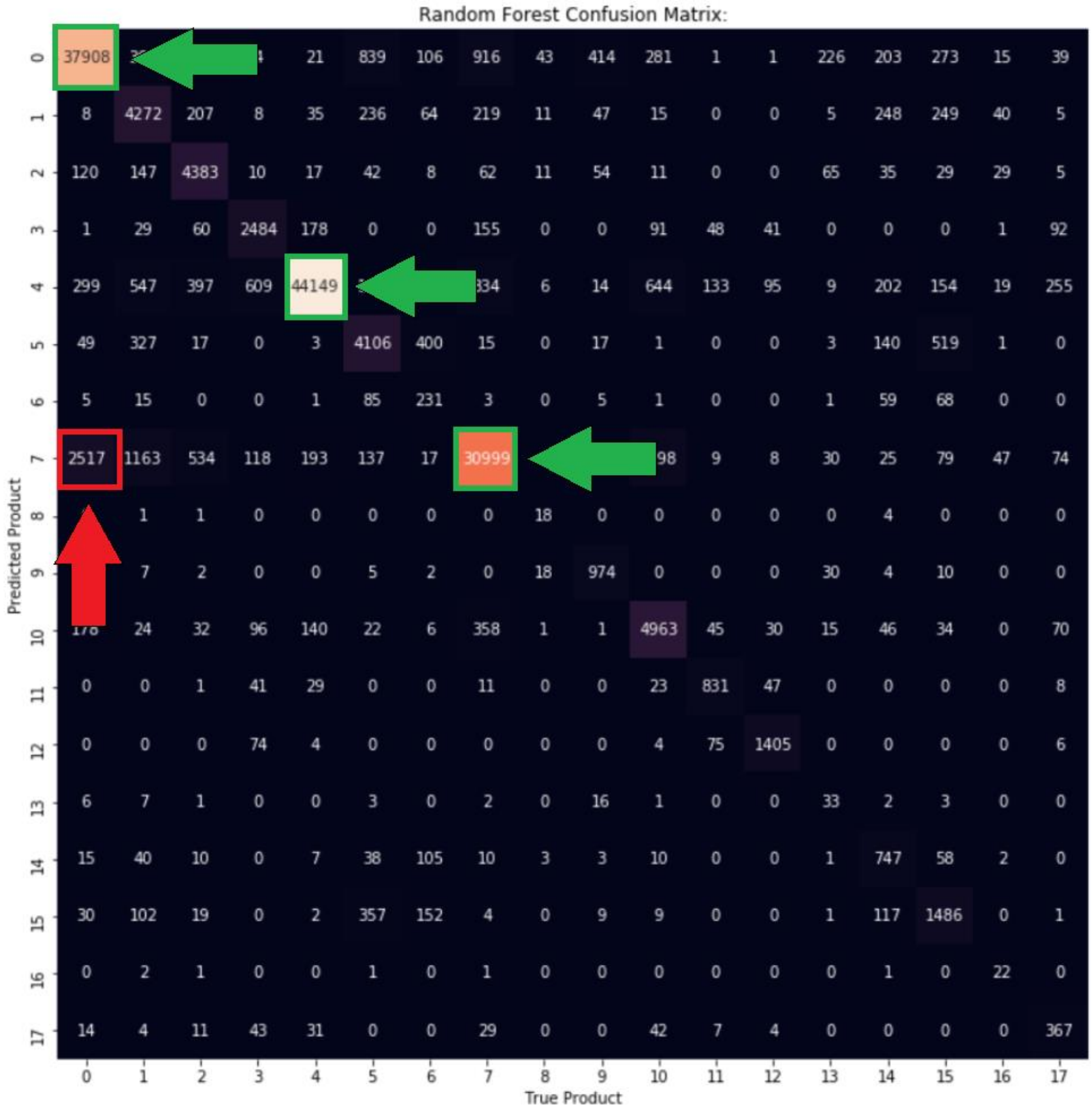




Random Forest

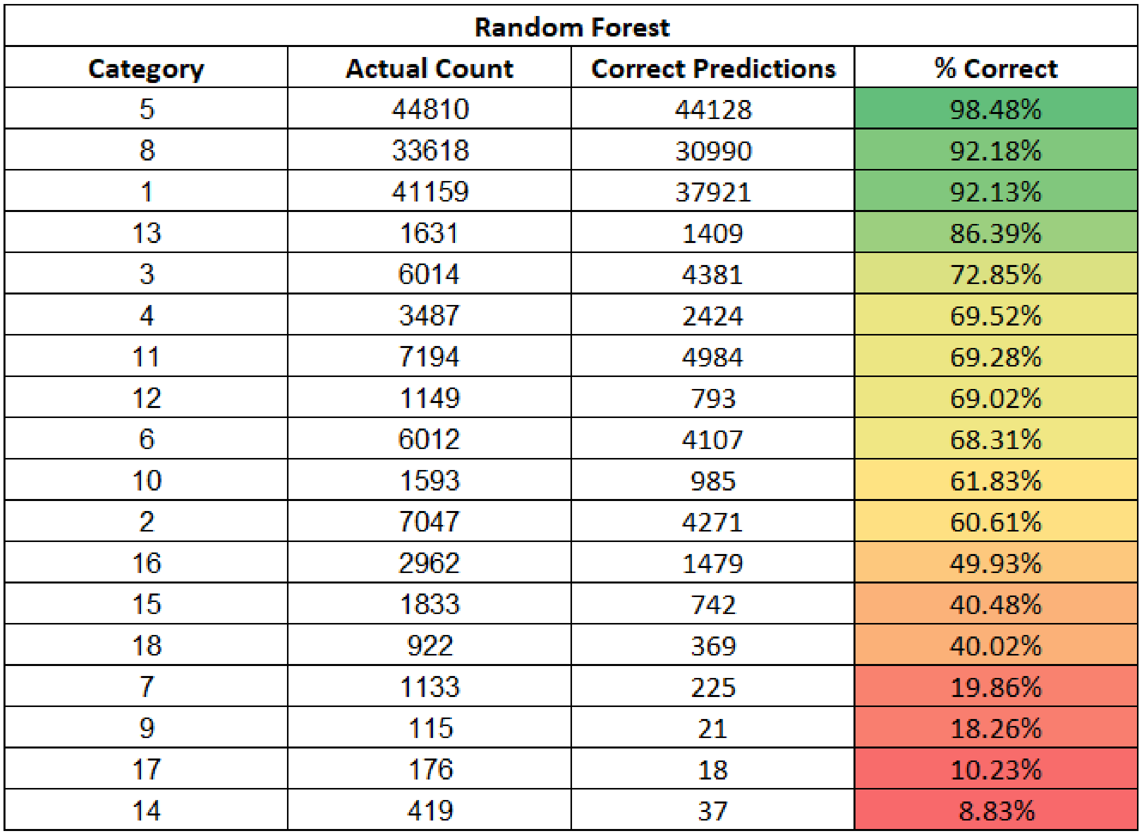
-Background: Random Forest is an ensemble learning method. It generates multiple decision trees. This makes it likely to be more accurate than a single tree, but computationally more expensive. It is also more difficult to interpret than a single tree.

-Findings: The Random Forest model accurately categorized 86% of the product categories. Most of the correctly categorized products were placed in category 1, 5, and 8 (corresponding with 0,4,7 below).


Random Forest Confusion Matrix:

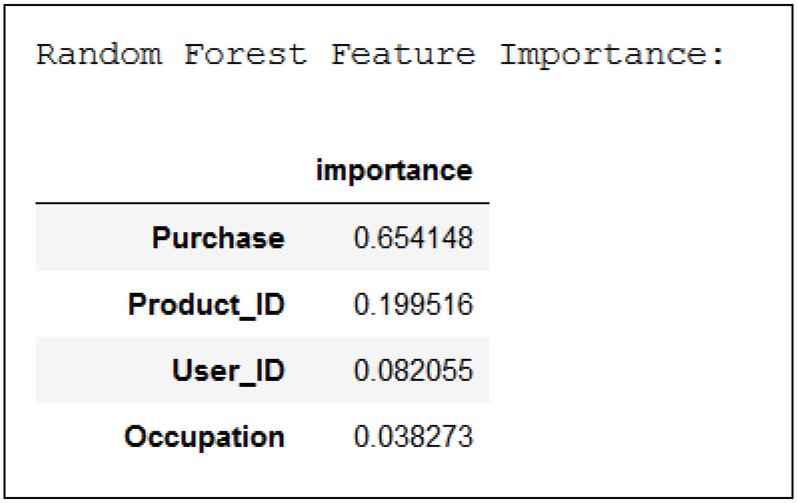
Predicted Product \ True Product	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	37908	[illegible]	[illegible]	[illegible]	21	839	106	916	43	414	281	1	1	226	203	273	15	39
1	8	4272	207	8	35	236	64	219	11	47	15	0	0	5	248	249	40	5
2	120	147	4383	10	17	42	8	62	11	54	11	0	0	65	35	29	29	5
3	1	29	60	2484	178	0	0	155	0	0	91	48	41	0	0	0	1	92
4	299	547	397	609	44149	[illegible]	[illegible]	334	6	14	644	133	95	9	202	154	19	255
5	49	327	17	0	3	4106	400	15	0	17	1	0	0	3	140	519	1	0
6	5	15	0	0	1	85	231	3	0	5	1	0	0	1	59	68	0	0
7	2517	1163	534	118	193	137	17	30999	[illegible]	[illegible]	98	9	8	30	25	79	47	74
8	[illegible]	1	1	0	0	0	0	0	18	0	0	0	0	0	4	0	0	0
9	[illegible]	7	2	0	0	5	2	0	18	974	0	0	0	30	4	10	0	0
10	178	24	32	96	140	22	6	358	1	1	4963	45	30	15	46	34	0	70
11	0	0	1	41	29	0	0	11	0	0	23	831	47	0	0	0	0	8
12	0	0	0	74	4	0	0	0	0	0	4	75	1405	0	0	0	0	6
13	6	7	1	0	0	3	0	2	0	16	1	0	0	33	2	3	0	0
14	15	40	10	0	7	38	105	10	3	3	10	0	0	1	747	58	2	0
15	30	102	19	0	2	357	152	4	0	9	9	0	0	1	117	1486	0	1
16	0	2	1	0	0	1	0	1	0	0	0	0	0	0	1	0	22	0
17	14	4	11	43	31	0	0	29	0	0	42	7	4	0	0	0	0	367

Random Forest did a better job predicting products in category 1, 5, and 8, getting 98% for category 5. This is possibly because there was more training data for these categories.



Random Forest			
Category	**Actual Count**	**Correct Predictions**	**% Correct**
5	44810	44128	98.48%
8	33618	30990	92.18%
1	41159	37921	92.13%
13	1631	1409	86.39%
3	6014	4381	72.85%
4	3487	2424	69.52%
11	7194	4984	69.28%
12	1149	793	69.02%
6	6012	4107	68.31%
10	1593	985	61.83%
2	7047	4271	60.61%
16	2962	1479	49.93%
15	1833	742	40.48%
18	922	369	40.02%
7	1133	225	19.86%
9	115	21	18.26%
17	176	18	10.23%
14	419	37	8.83%

The most important feature was Purchase amount.



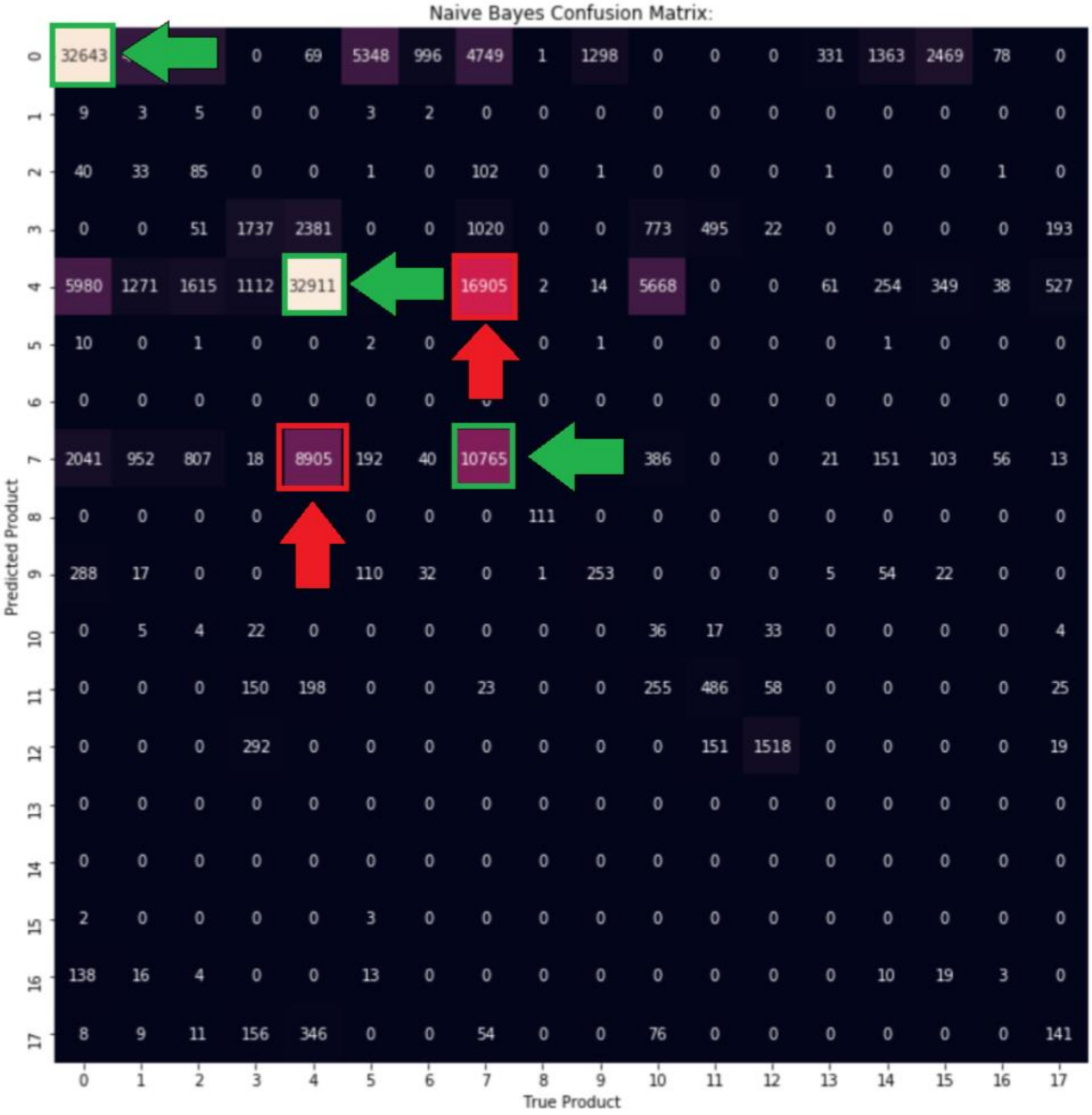
```
Random Forest Feature Importance:
```

	importance
Purchase	0.654148
Product_ID	0.199516
User_ID	0.082055
Occupation	0.038273

Naive Bayes

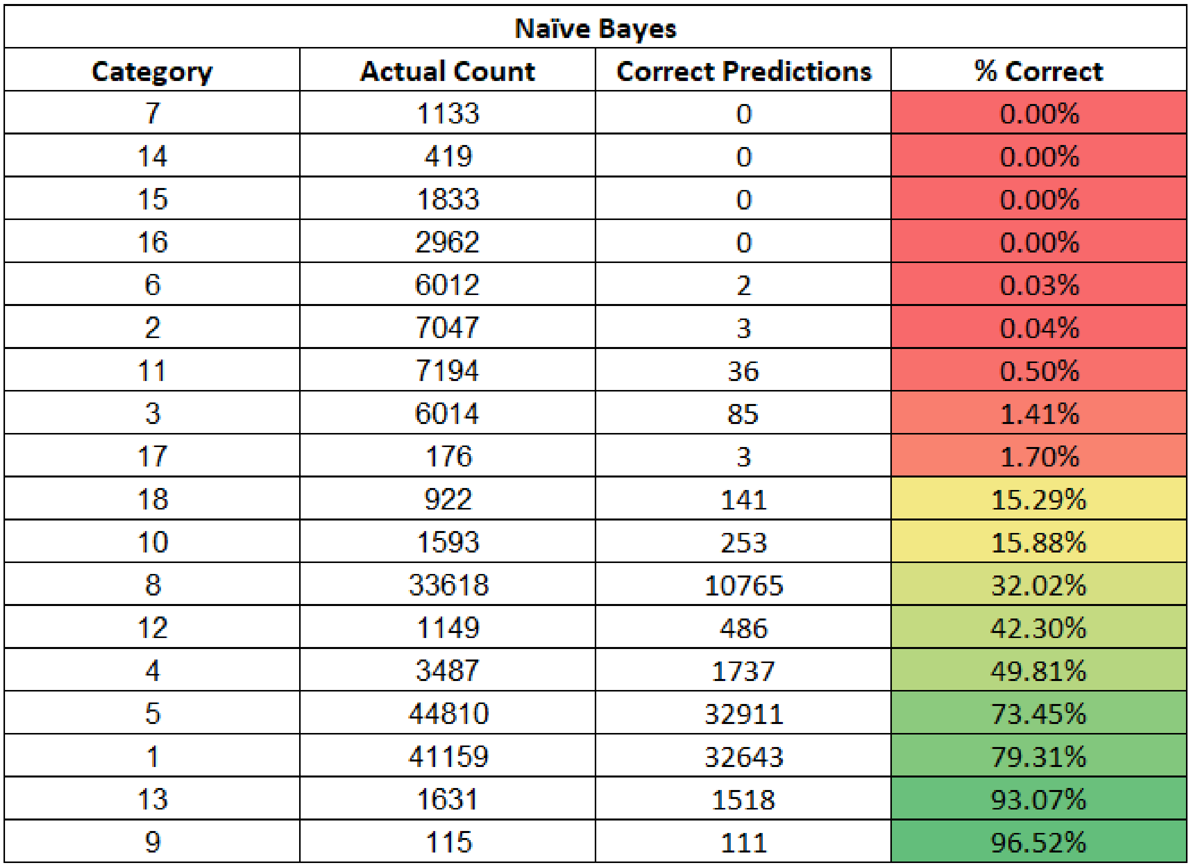
-Background: Naive Bayes is computationally inexpensive and works well on data even if dimensionality is high. However, it treats all events as if they were independent.

-Findings: The Naive Bayes model only accurately categorized 50% of the product categories. Most of the correctly categorized products were placed in category 1, 5, and 8 (corresponding with 0,4,7 below). The largest wrong prediction was 16,905 instances predicted as 5 that were actually 8 (in bright pink/red below). The second largest wrong prediction was 8,905 instances predicted as 8 that were actually 5.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	32643	[illegible]	[illegible]	0	69	5348	996	4749	1	1298	0	0	0	331	1363	2469	78	0
1	9	3	5	0	0	3	2	0	0	0	0	0	0	0	0	0	0	0
2	40	33	85	0	0	1	0	102	0	1	0	0	0	1	0	0	1	0
3	0	0	51	1737	2381	0	0	1020	0	0	773	495	22	0	0	0	0	193
4	5980	1271	1615	1112	32911	[illegible]	[illegible]	16905	2	14	5668	0	0	61	254	349	38	527
5	10	0	1	0	0	2	0	[illegible]	0	1	0	0	0	0	1	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	2041	952	807	18	8905	192	40	10765	[illegible]	[illegible]	386	0	0	21	151	103	56	13
8	0	0	0	0	[illegible]	0	0	0	111	0	0	0	0	0	0	0	0	0
9	288	17	0	0	[illegible]	110	32	0	1	253	0	0	0	5	54	22	0	0
10	0	5	4	22	0	0	0	0	0	0	36	17	33	0	0	0	0	4
11	0	0	0	150	198	0	0	23	0	0	255	486	58	0	0	0	0	25
12	0	0	0	292	0	0	0	0	0	0	0	151	1518	0	0	0	0	19
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
16	138	16	4	0	0	13	0	0	0	0	0	0	0	0	10	19	3	0
17	8	9	11	156	346	0	0	54	0	0	76	0	0	0	0	0	0	141

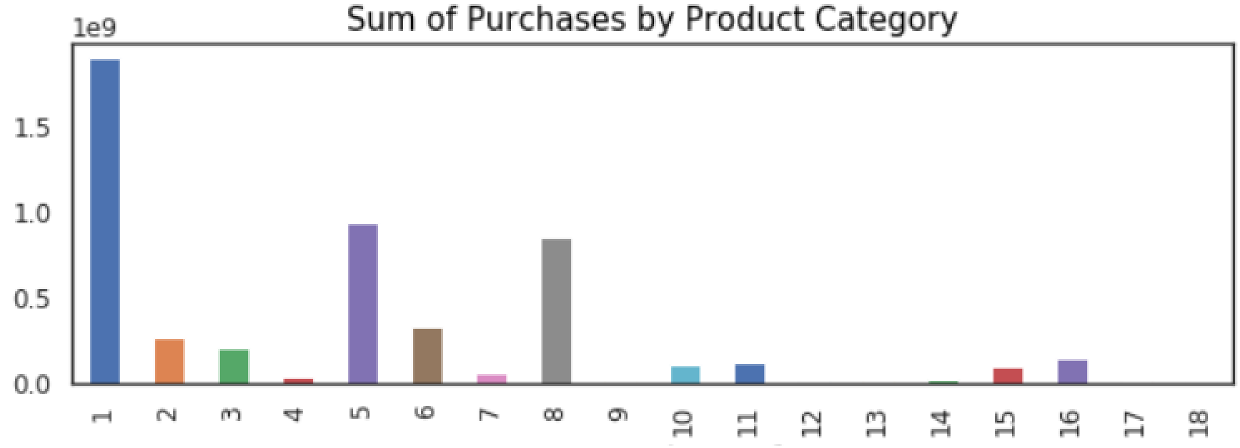
Naive Bayes did a better job predicting products in category 1, 9, and 13, getting 97% for category 9. However, for half of the product categories, it correctly categorized less than 2%. Three categories had no correct predictions at all.



Naïve Bayes			
Category	**Actual Count**	**Correct Predictions**	**% Correct**
7	1133	0	0.00%
14	419	0	0.00%
15	1833	0	0.00%
16	2962	0	0.00%
6	6012	2	0.03%
2	7047	3	0.04%
11	7194	36	0.50%
3	6014	85	1.41%
17	176	3	1.70%
18	922	141	15.29%
10	1593	253	15.88%
8	33618	10765	32.02%
12	1149	486	42.30%
4	3487	1737	49.81%
5	44810	32911	73.45%
1	41159	32643	79.31%
13	1631	1518	93.07%
9	115	111	96.52%

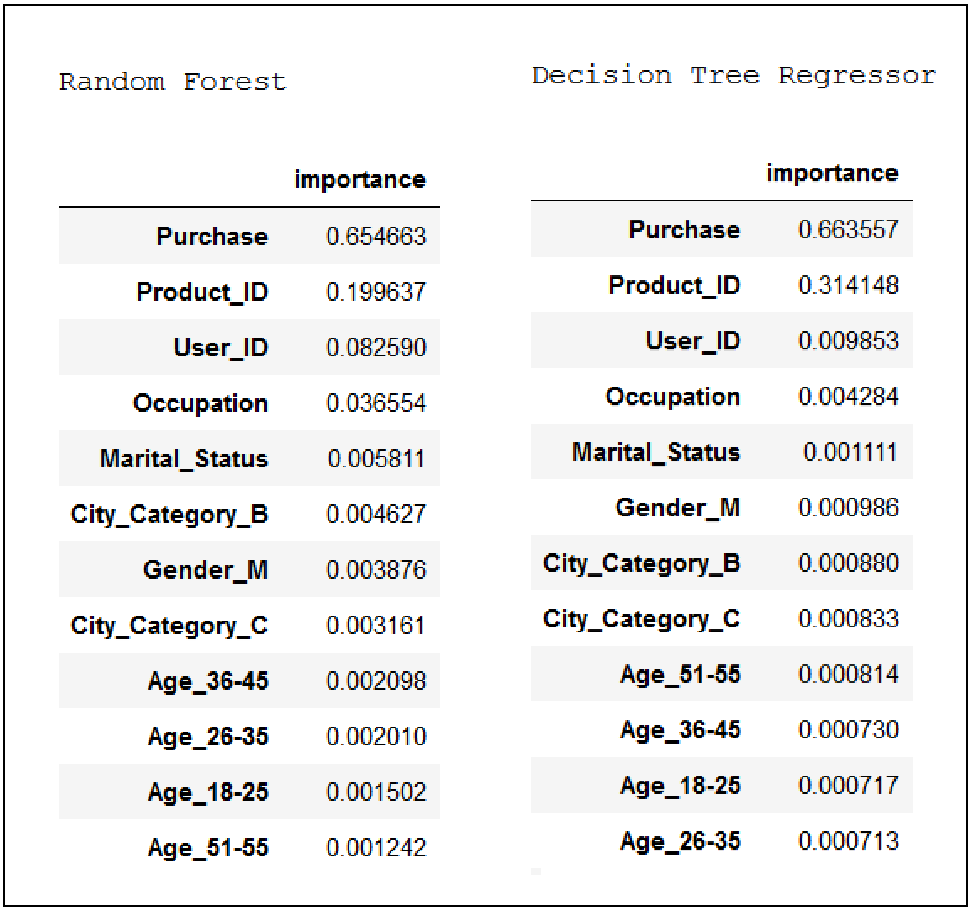
The largest numbers of the correctly categorized products were placed in category 1, 5, and 8 (corresponding with 0,4,7 on the confusion matrix). However, the Product Category 8 (7 in the confusion matrix) gives us more insight. The Naive Bayes model assumes that the majority of the products fall into one of these three categorizes, but it doesn't necessarily know which one. Taking a look at the distribution of products in each category allows for further understanding.

The product category chart (on the next page) depicts that most of the products do fall into one of these three categories. Consequently, our model assumes the same. From the Naive Bayes model, it is evident that our model has an easier time categorizing products into category 1 than with products that are in category 8 (7 in the confusion matrix). In contrast, the Random Forest model had an easier time with category 5 and a harder time with category 8 and 1.



Other Models:

A single decision tree was run, and surprisingly, it performed better than Random Forest (91% vs 86%). Feature importance was similar for both models, but the decision tree prioritized gender over city category.



Random Forest

	importance
Purchase	0.654663
Product_ID	0.199637
User_ID	0.082590
Occupation	0.036554
Marital_Status	0.005811
City_Category_B	0.004627
Gender_M	0.003876
City_Category_C	0.003161
Age_36-45	0.002098
Age_26-35	0.002010
Age_18-25	0.001502
Age_51-55	0.001242

Decision Tree Regressor

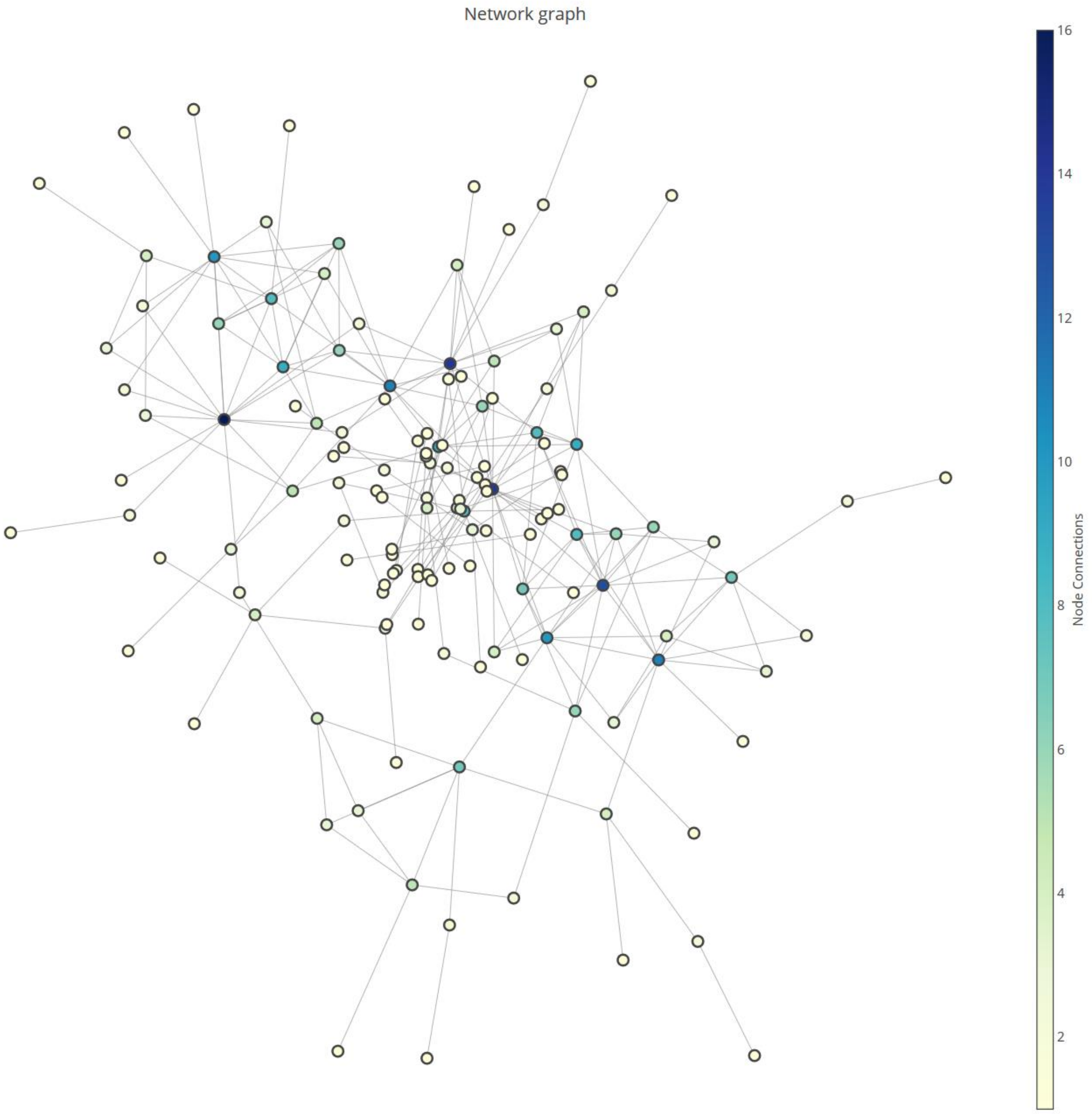
	importance
Purchase	0.663557
Product_ID	0.314148
User_ID	0.009853
Occupation	0.004284
Marital_Status	0.001111
Gender_M	0.000986
City_Category_B	0.000880
City_Category_C	0.000833
Age_51-55	0.000814
Age_36-45	0.000730
Age_18-25	0.000717
Age_26-35	0.000713

KNN Regressor was run to try to predict product category, and gave 62% accuracy.

A Linear Regression was run on product category, but this only had a 10% accuracy. This makes sense, because linear regression works better for continuous variables, and product category is discrete.

K-Means Clustering was attempted on customer data but only gave 25% accuracy for City Category and 8% accuracy for Occupation category.

Association Rules was run and visualized, but it was difficult to interpret. The largest number of connections any one node had was 16.



Model Summary:

The Random Forest algorithm classified the product category of purchase with an 86% precision. This was surpassed only by the Decision Tree model which generated an impressive 91% accuracy. The KNN Regressor and Naive Bayes models were able to classify the product category with 62% and 50% accuracy. K-Means Clustering gave results of 25% and 8% accuracy. These results demonstrate the Decision Tree model is the best for predicting the purchasing behavior of customers. This can be an advantage to stores preparing for the Black Friday mayhem that occurs every year.

Recommendations

When it came to shopping for items in particular categories, all variables being analyzed had similar spending habits. The best thing to do in this situation would be to prioritize the products that were most profitable. This happened to be Product Category 1 (38%), 5 (18%), and 8 (17%).

As stated above, the Decision Tree model was found to be the most valuable and may prove valuable for future use.

Further work could include investigating why customers from city category C aren't spending more and trying to do more marketing in that city. Additionally, our client can either embrace the fact that single males aged 26-35 are their main customers and market to them, or try to determine if there are marketing campaigns that might bring in more customers who are married, female, and in different age groups.

References

https://datahack.analyticsvidhya.com/contest/black-friday/

https://www.kaggle.com/mehdidag/black-friday

https://www.kaggle.com/sriharshaatyam/eda-and-predictions-using-various-ml-models

https://stackoverflow.com/questions/332289/how-do-you-change-the-size-of-figures-drawn-with-matplotlib

https://plot.ly/python/network-graphs/

https://towardsdatascience.com/running-random-forests-inspect-the-feature-importances-with-this-code-2b00dd72b92e