Predicting the Gender of the Consumer using Black Friday Dataset

Caixuan Sun

1. Introduction

Retail data has been increasing exponentially in both volume and value. With the massive amounts of data available, the retail industry is able to rely on data science to extract insights to attract and retain customers, discover trends, operate business more efficiently and ultimately increase sales and reduce costs. There are several angles to work on such as customer experience, marketing, supply chain logistics and merchandising. For the first project, I would like to focus on consumers and their behaviors. It is greatly beneficial for the retail store to know their customers well. By identifying who the retail store is selling to and what the customers are buying, better decisions regarding personalized offers and product recommendations could be made. The specific goal for the project is to predict the gender of the consumer which could help the retail store to construct more targeted marketing strategies.

Data is obtained from Kaggle: https://www.kaggle.com/mehdidag/black-friday. The dataset is a sample of transactions made in a retail store for a period of one month. It contains customer demographics such as age, gender, occupation and marital status, product details including product id and product category, and purchase amount of each product by each consumer for the whole month.

There are mainly three parts in this project. First is to collect and clean data. Second is to do some exploratory data analysis. Last is to build machine learning models to solve for the task proposed above and then compare the performance of different models.

2. Data Wrangling

2.1 Data Description

The data set consists of 537,577 rows and 12 columns. Each row represents a transaction record of one particular product bought by one consumer during this month. Each customer may have several rows to record their purchases of different products and purchase amount of that product in this month. 5,891 customers have purchased products and 3,623 different products were sold at this store during this month.

2.2 Data Cleaning

Data is already pretty clean. I've performed several basic check ups as followings. To begin with, I checked if there is any duplicated rows. Then I converted all column names into lower case and examined if all column labels are of type string. In this dataset, all features except purchase are categorical variables. Only purchase is continuous numeric variables. Thus I converted all categorical features into data type of category to save memory and improve performance.

Next step is to check missing values. Only two features, product_category_2 and product_category_3 have null entries. Also there are no other form of missing values found. After further examination of the three features for product categories, we find that product_category_2 and product_category_3 might be sub categories. Missing values may be due to the lack of sub categories for certain products. All

products fall into one and only one main category of product_category_1. Some of them may have sub categories, some may not. I replaced nulls with value 0.

I did a boxplot for feature purchase, the only one numeric feature. According to the boxplot, there are many outliers. I used 1.5 times IQR (interquartile range) as a threshold to extract those outliers to do further check up. There are 2665 outliers identified. My opinion is that some products are quite expensive comparing with others. Many customers may have waited until Black Friday to purchase those products. I will just leave those observations there.

Last I aggregated the granular data into customer-level which has one row representing one unique customer. Columns include all demographics, and two new aggregated features: total purchase of each customer and total number of unique products purchased during this month.

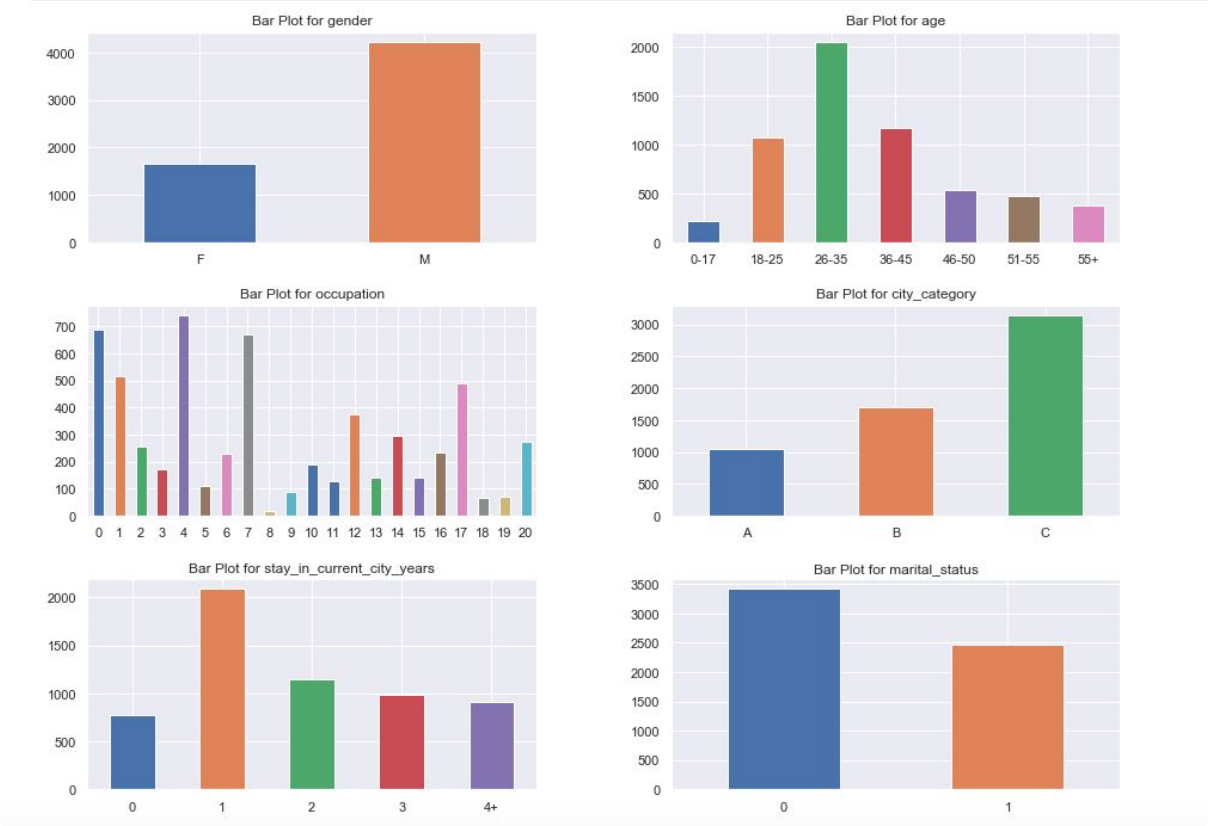
3. EDA

3.1 Univariate Analysis

3.1.1 Distribution of All Demographic Features

First I explored all demographic feature by plotting a bar graph for each of them.

Figure 1. Bar Plot for All Demographic Features



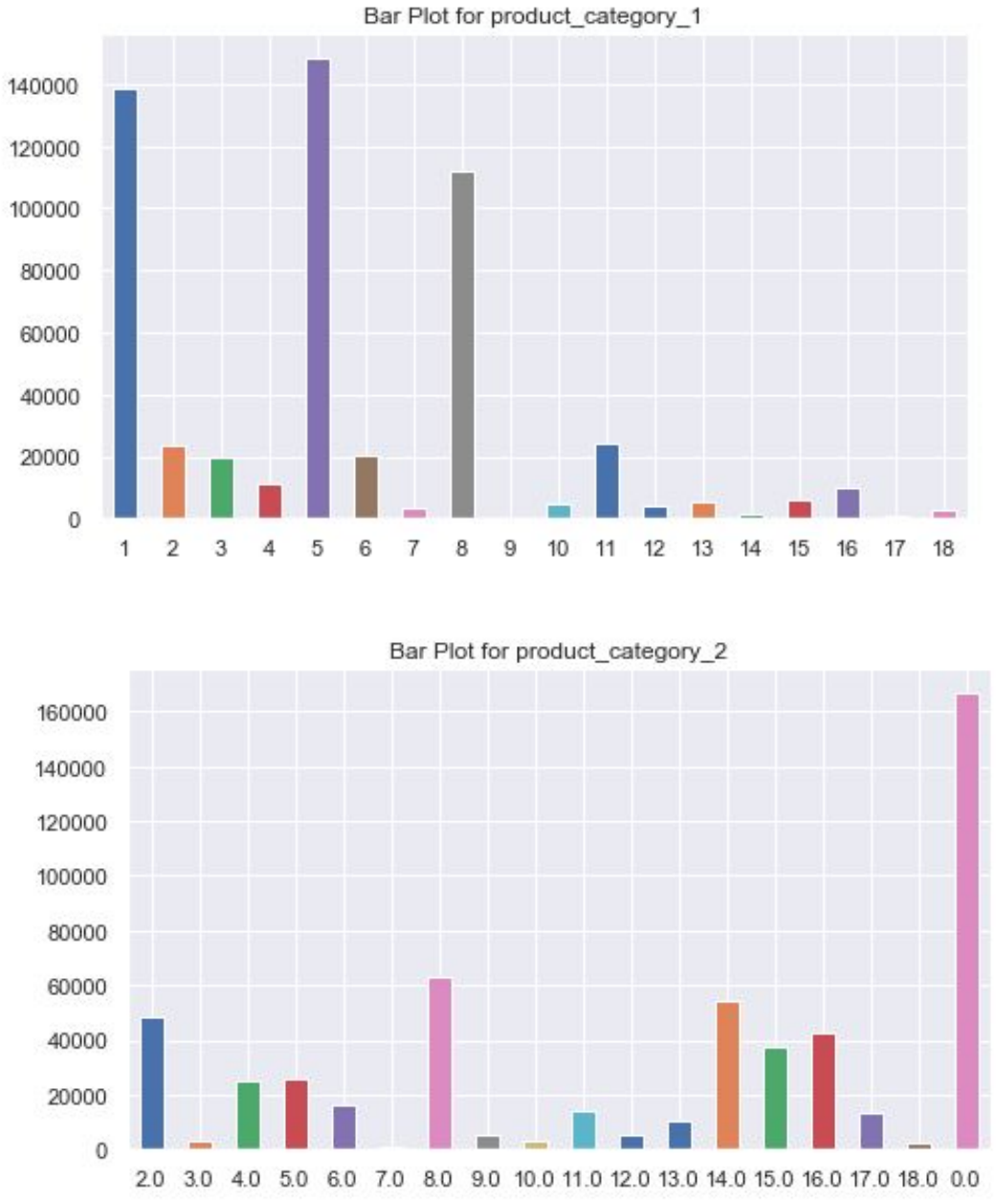
From the bar graph, we find that there are more male customers than female customers. We will further investigate gender column later since it's our target variable.

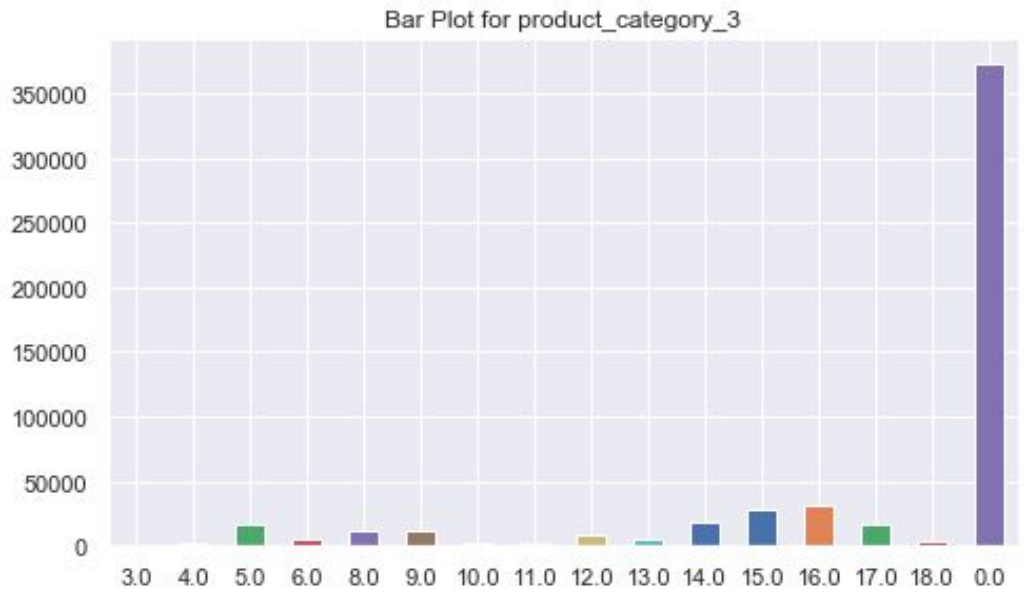
Out of 5891 customers, there are more than 2000 between 26 and 35 years old. Most purchases are made by customers between 18 and 45 years old. There are 20 categories for occupation. A majority of customers are from categories 0, 1, 4, 7 and 17. More than half of the customers are from city category C. It might be due to the location of the store. Half of the customers to this store are new to the current city. There is a decreasing trend. The longer the customers stay in current city, the less they purchase at this store. The reason maybe that new people need to buy more stuff for their home. Or because that those who live there longer know a better place to buy things. This store may need to pay attention to this problem. Single customers

are more than married customers in this store. A simple conclusion is that this store is popular for male, people living in city C, people between the age of 26 and 35, new to their current city and single persons.

3.1.2 Distribution of Product Categories

Figure 2. Bar Plot for Product_Categories





From the bar plot of product_category_1, product_category_2, and product_category_3, we find that category 1, 5, 8 are the most popular categories based on product_category_1. There are a lot of null values for the other two categories.

3.1.3 Numerical Features

Table 1. Statistic Summary of Numerical Features

	total_purchase	n_purchases
count	5.891000e+03	5891.000000
mean	8.517515e+05	91.253947
std	9.329978e+05	105.929800
min	4.410800e+04	5.000000
25%	2.349140e+05	25.000000
50%	5.126120e+05	53.000000
75%	1.099005e+06	114.000000
max	1.053678e+07	1025.000000

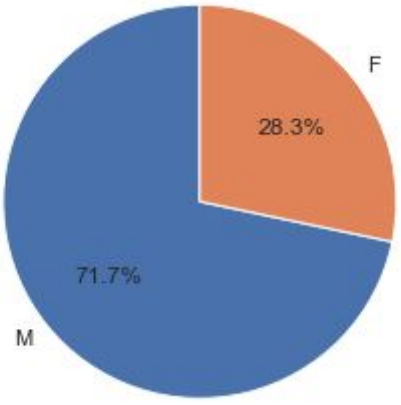
From the summary statistics, we can find that both of them are quite right skewed. Besides, they're not at the same scale, standard deviation of total_purchase is large. We have to deal with this situation later in order to get a better fit from machine learning model.

3.2 Bivariate Analysis

Since our task is to predict the gender of the customer, 'gender' is our target variable. First we'll focus on visualizing and exploring the relationship between 'gender' and all the other candidates for predictors.

3.2.1 More Analysis on Gender Itself

Figure 3. Gender Distribution

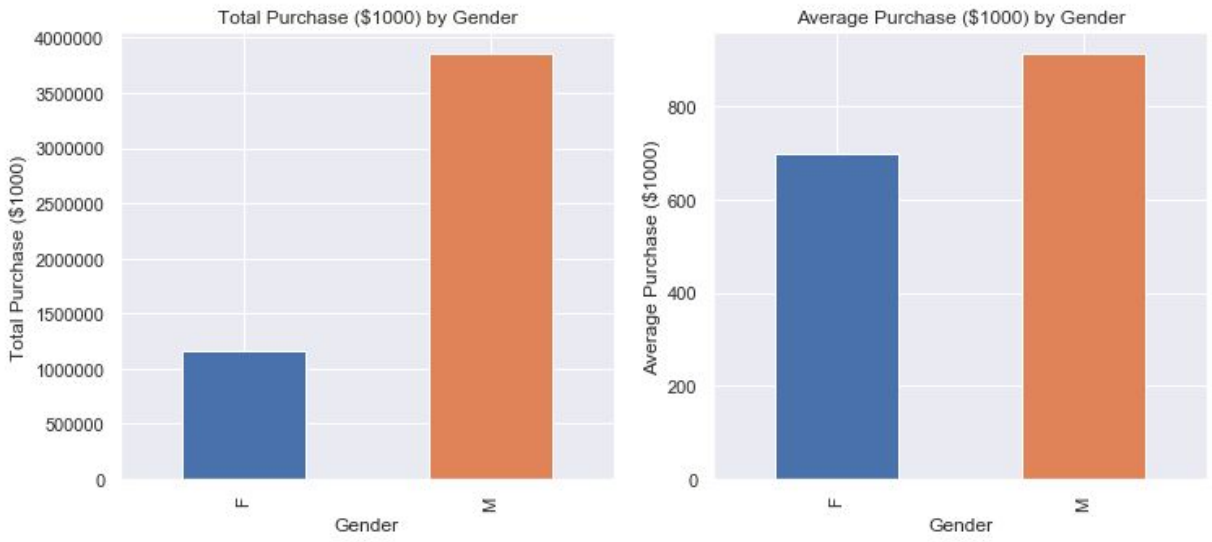


There are more male customers than female customers at this store. Specifically, 71.7% of all customers at this store are male, the rest 28.3% are female.

3.2.2 Gender and Purchase

We know the customers' gender distribution at this store, how about their purchase power respectively?

Figure 4. Total Purchase (in $1000) and Average Purchase (in $1000) across Gender



By comparing female and male customers' total purchase and average purchase amount, we see that male customers are spending much more than female customers cause there are much more male consumers at this store. While in terms of average purchase amount, the difference decreases a lot.

Table 2. Summary of Purchase by Gender

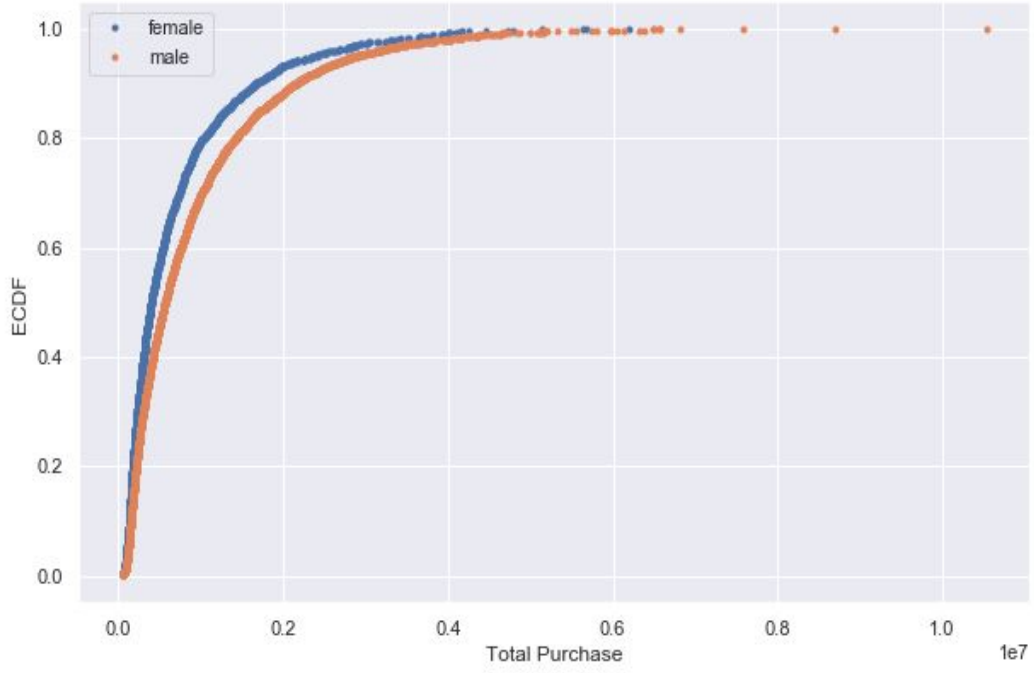
	count	gender_proportion (%)	total_purchase	purchase_proportion (%)	average_purchase
F	1666	28.280428	1164624021	23.210462	699054.034214
M	4225	71.719572	3853044357	76.789538	911963.161420

Total purchase by female customers is 1.16 billion, total purchase by male customers is 3.85 billion. Female accounts for 28.3% of total population at this store,

while accounts for 23.2% of total purchase. Male accounts for 71.7% of total population and 76.8% of total purchase at this store.

Statistically speaking, are female and male purchase drawn from the distribution? Do they have the same mean purchase amount? Let's first check the empirical cumulative distribution for female and male total purchase.

Figure 5. ECDF of Total Purchase by Female and Male Customers



It seems that the two CDFs are quite separated. ThenI conduct a two-sided test for the null hypothesis that two samples of female purchase and male purchase are drawn from the same distribution. P-value is extremely small, thus we reject the null that they have the same distribution. I also conduct a two-sided test for the null hypothesis that mean purchase for male and female customers are the same.

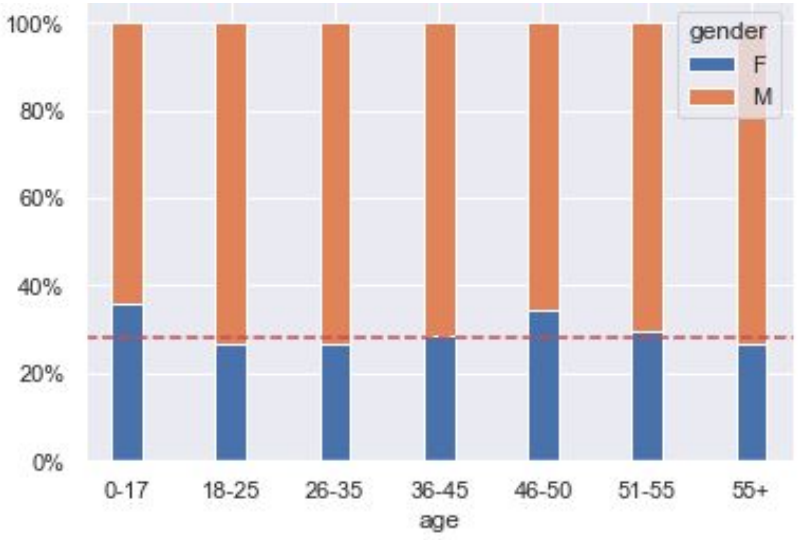
Next, I conducted the same two tests for number of purchases. Both have extremely small p-values. Therefore, we conclude that difference between number of purchases by female and male are statistically significant.

3.2.3 Gender and Age

In this subsection, I will check how gender distributed across different age groups.

Figure 6. Gender Distribution Across Different Age Groups

Figure 7. Percentage View of Gender Distribution Across Different Age

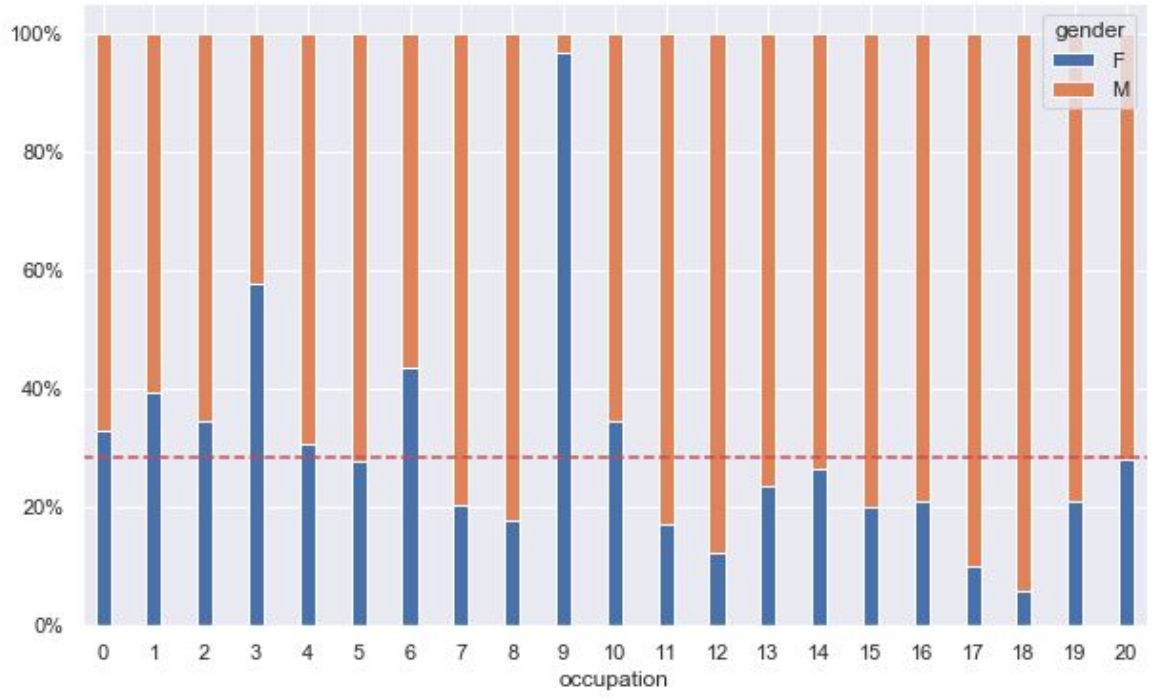


From the second plot, we can see clearly how gender distribution in each age group is different from its population distribution in this store. I added a horizontal line at 28.3% which is the proportion of female customers in this store. There are relatively

more female customers for age group 0-17 and 46-50. There's slightly less female in age group 18-25, 26-35 and 55+.

3.2.4 Gender and Occupation

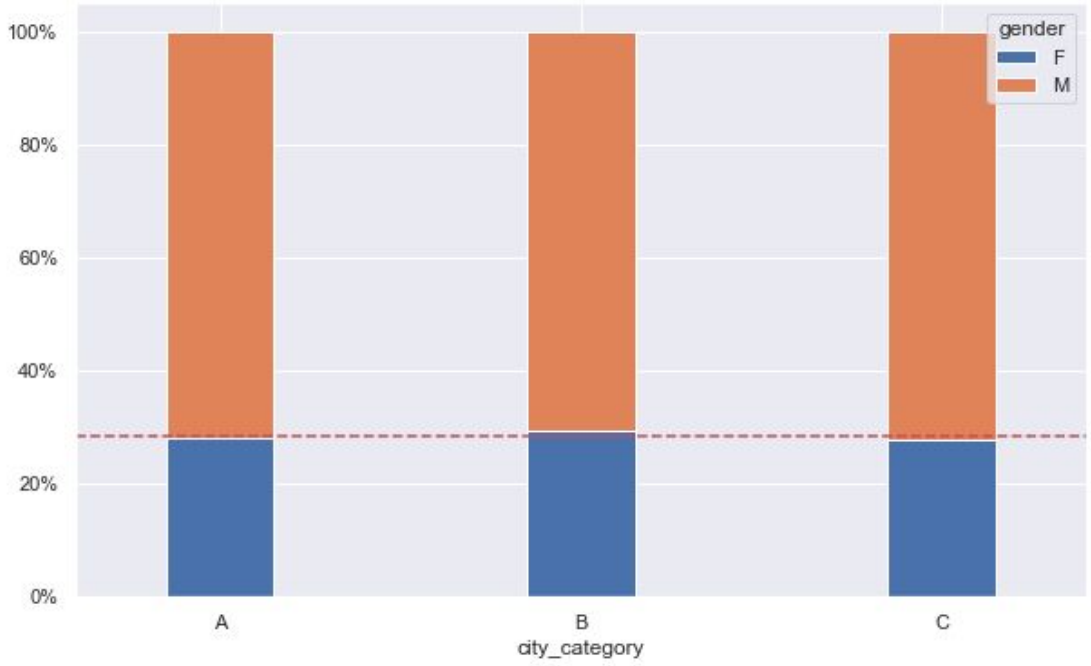
Figure 8. Percentage View of Gender Distribution Across Occupation



Most of the customers with occupation category 9 are female. Occupation 17 and 18 are dominated by male. Gender distribution for categories such as 4, 5, 14, and 20 are pretty the same as the whole sample distribution of gender at this store.

3.2.5 Gender and City Category

Figure 9. Percentage View of Gender Distribution Across City

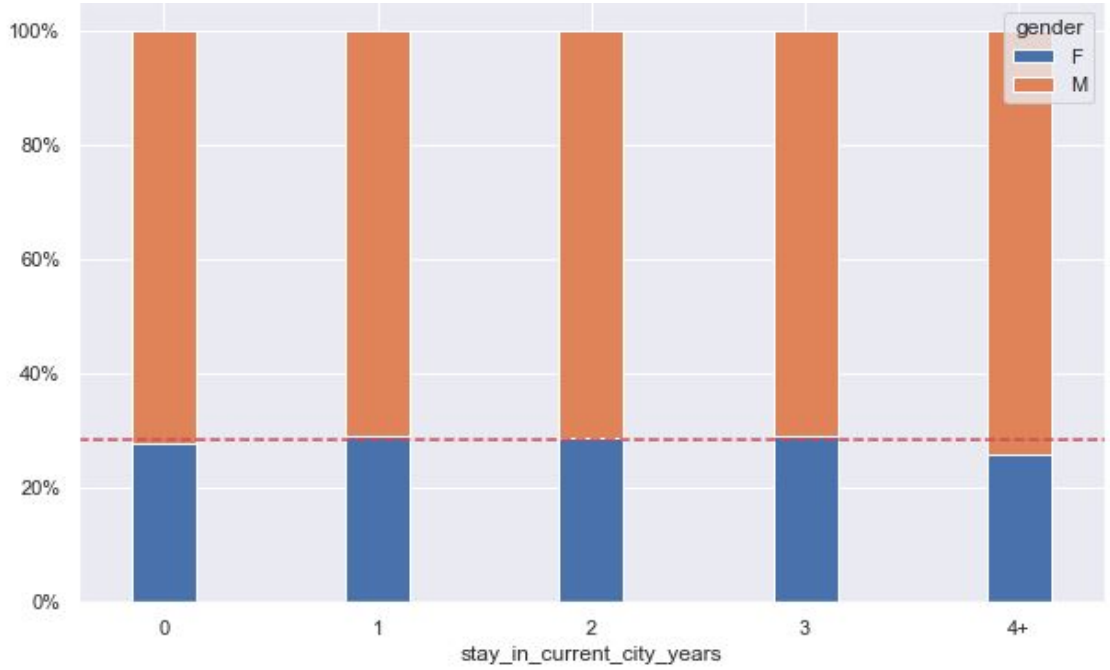


Gender distribution for each city category is nearly the same as its population distribution in this store. This feature ‘city_category’ may not have much predictive power for gender.

3.2.6 Gender and Stay in Current City Years

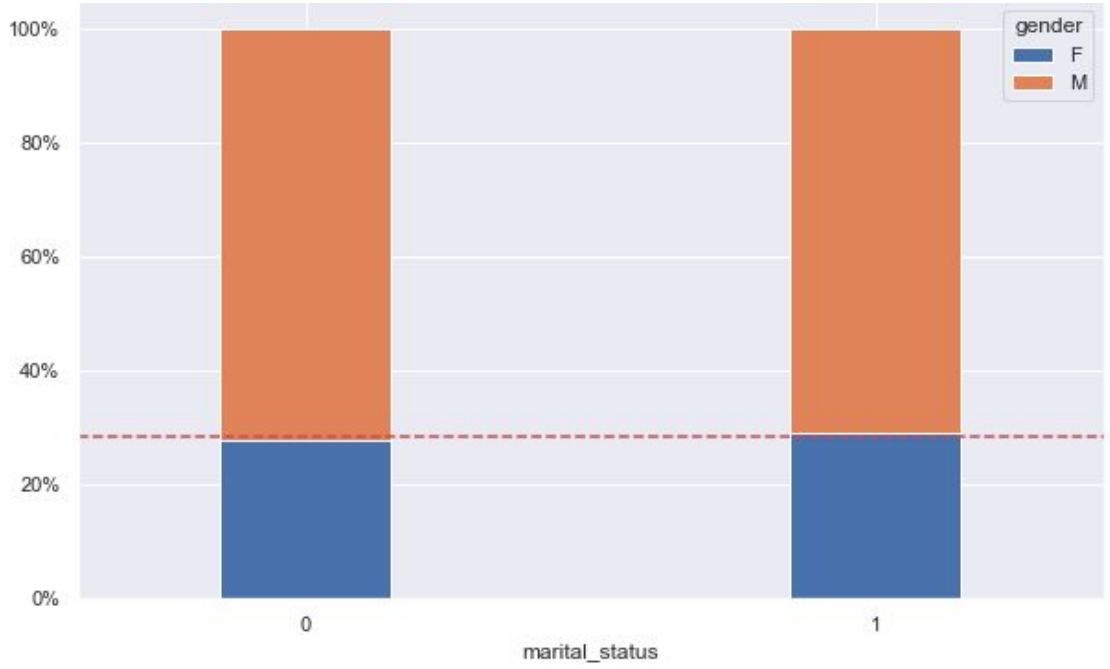
From the figure below, we can see that gender distribution in each category for stay_in_current_city_years feature is nearly the same as the whole gender distribution in this store except that for category 4+ there are slightly more male proportionately speaking.

Figure 10. Percentage View of Gender Distribution Across Different Stay_in_Current_City_years



3.2.7 Gender and Marital Status

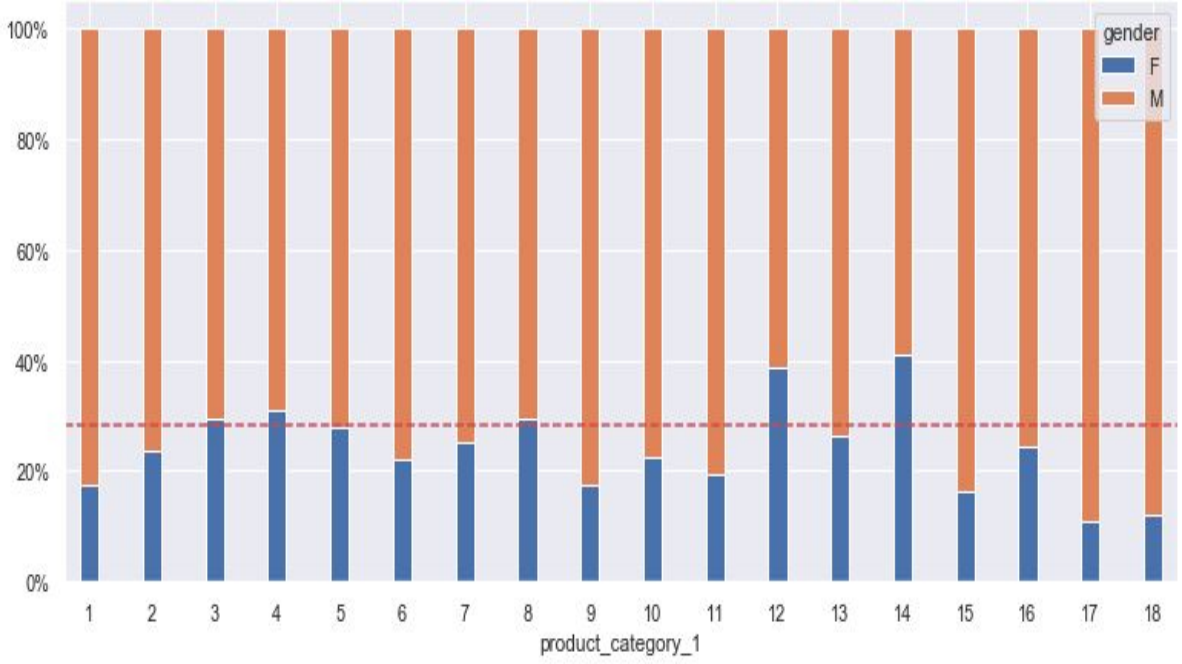
Figure 11. Percentage View of Gender Distribution Across Different Marital Status



Gender distribution are pretty the same for single and married group, and the gender distribution of this store.

3.2.8 Gender and Product Category 1

Figure 12. Percentage View of Gender Distribution Across Different Product Categories

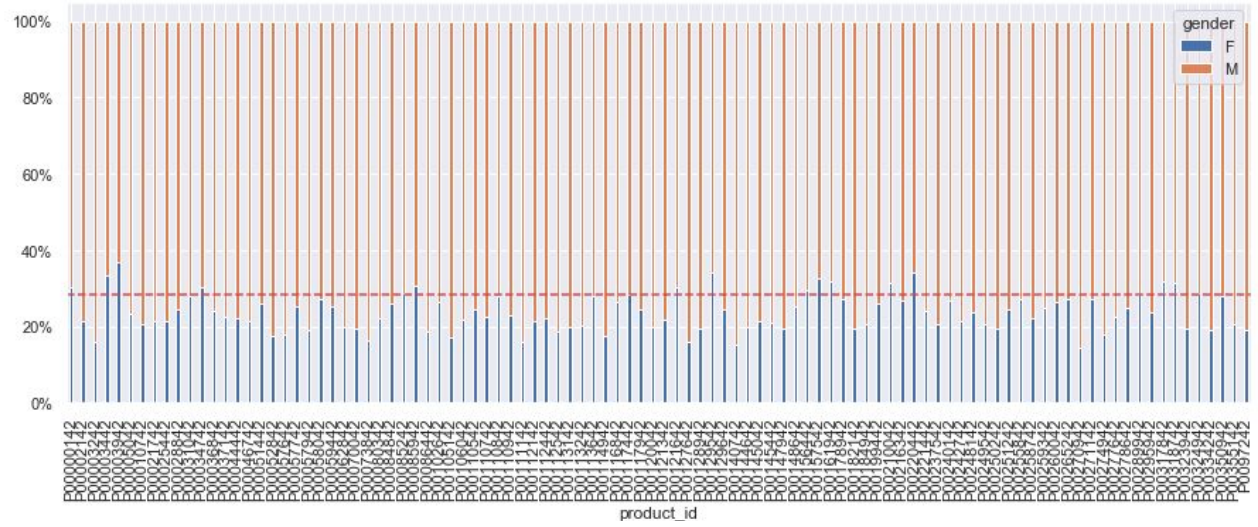


For categories 3, 4, 5, 8, and 13, the distribution of gender is nearly the same as the distribution at this store. For categories 1, 9, 12, 14, 15, 17, 18, proportion of female and male customers are deviating from the store's gender proportion. These categories may be important features when predicting gender.

3.2.9 Gender and Products

The top seller was purchases by 1858 customers, and many of the products were bought by only one of the customers. We can also see product_id column is useful in predicting gender since female and male are buying different products.

Figure 13. Percentage View of Gender Distribution Across Top 100 Products



3.3 Other Bivariate Analysis

One more interesting thing is the relationship between purchase and city category.

Figure 14. Pie Chart for City Category

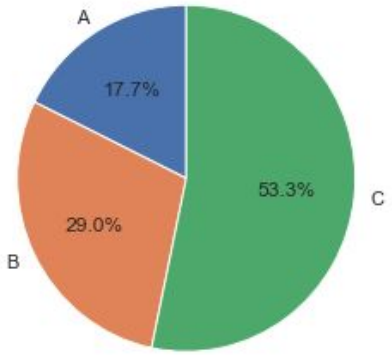
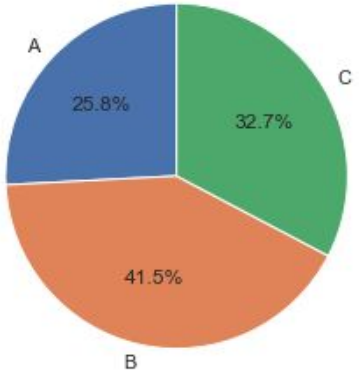


Figure 15. Total Purchase by Each City Category



Although number of consumers in city B are in the middle, accounting for 29% of all consumers of this store, their purchasing power is the highest among all three city categories, around 41.5% of total purchase amount is bought by customers from city category B.

4. Machine Learning

4.1 Introduction

Our task here is a supervised learning problem. We are given labeled data, e.g. we already know the gender of each customer. Classification algorithms will be used to classify the gender or customers. We have two classes, female (0) and male (1).

Some classification algorithms to be explored:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Random Forest Classifier
- Gradient Boosting Classifier

Evaluation metrics to be used:

- Accuracy score
- Metrics from confusion matrix and classification report: recall, precision, f1 score
- ROC curve and Area under ROC curve

4.2 Data Preprocessing

4.2.1 Standardizing Numerical Data

As we find from the summary statistic of the two numerical features, total_purchase and number_of_purchases, they are on different scales and total_purchase has very high variance. Standardization has to be applied for the linear based model to perform better. I chose to use log transformation.

4.2.2 Encoding Categorical Data

Before encoding all the categorical features, I have to first decide what products to be included in our feature set. I will select several products which have greater deviation from the general gender distribution. I chose a threshold of 5% to extract the products of which more than 28.8% + 5% or less than 28.8% - 5% were bought by female customers. As a result, 11 products from the top 100 sellers are included and converted to dummies. Another nominal categorical variable, ‘occupation’, is also converted to dummy variables.

Besides of nominal categorical variables, we also have ordinal ones such as ‘age’ and ‘stay_in_current_city_years’, I used the LabelEncoder() function to encode the labels of these two features.

After preprocessing data, we have 5891 rows for 5891 unique customers and 54 features.

4.2.3 Model Building

For the five models tried in this project, I first train them with default settings and then tune hyperparameters to improve the model performance. Our data is first split into a training set and a test set. We then use the training set to do gridsearch and 5 fold cross validation to pick the optimal parameters. Next we train the model using the obtained optimal parameters on the whole training set and then test on the test set.

4.2.3.1 Logistic Regression

I implement Logistic Regression using the sklearn's LogisticRegression class. I tuned two hyper parameters:

- C: inverse of regularization strength with smaller values specirying strong regularization.
- penalty: 'l1' or'l2'.

The testing accuracy is 0.7746. The performance of the model is further described in the confusion matrix and ROC curve below.

Figure 16. Confusion Matrix Using Logistic Regression

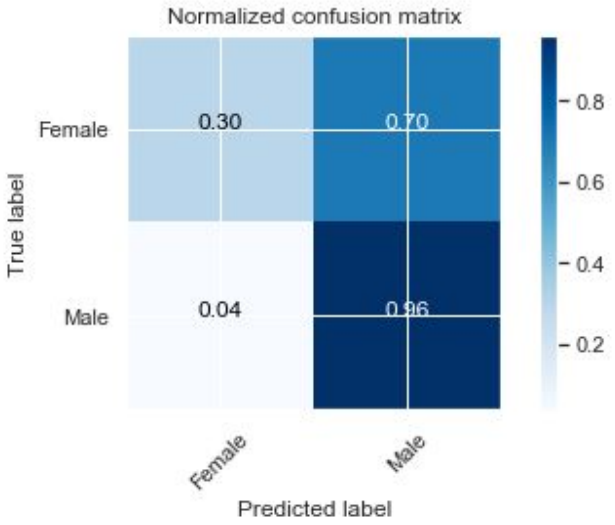


Figure 17. ROC Curve Using Logistic Regression

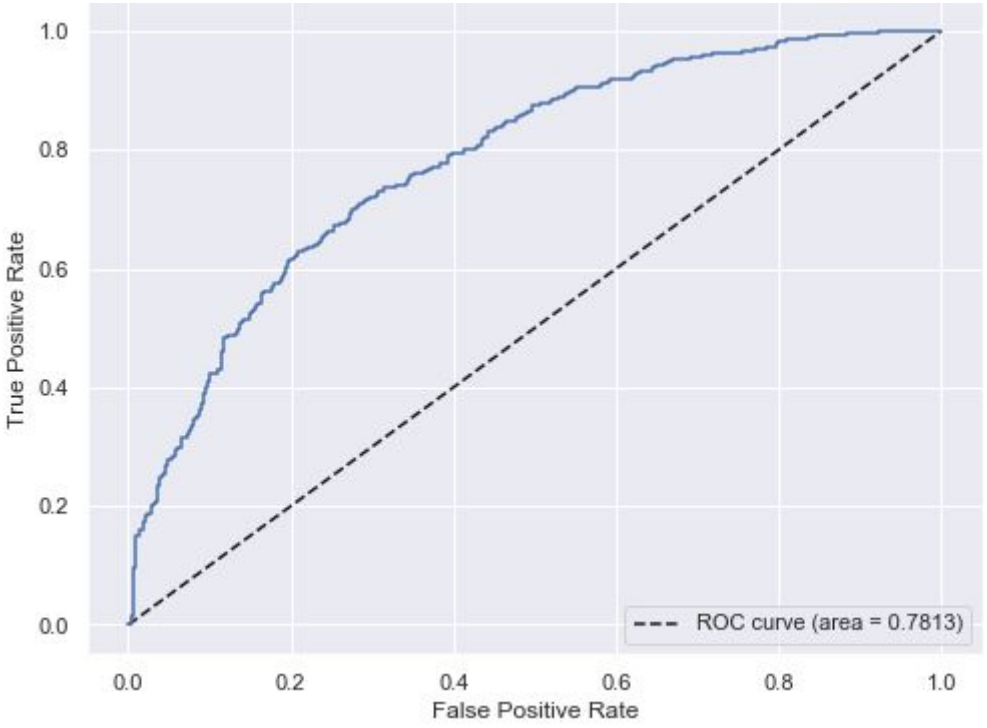


Table 3. Coefficients of Logistic Regression

product_category_1_1	0.024	product_category_1_12	-0.170	total_purchase_log	0.208
product_category_1_8	-0.02	product_category_1_14	-0.358	product_category_1_5	0.002
product_category_1_18	0.506	product_category_1_15	0.150	product_P00270942	0.264
occupation_9	-2.920	occupation_3	-1.216	occupation_17	0.609

I also checked the logistic regression coefficients the sign of which indicates the direction of the relationship between a predictor and the target variable ‘gender’. Some of them are listed above. You can find the full list at https://github.com/caixuansun/Springboard/blob/master/capstone_project_1/capston

e_project_1_final.ipynb. I find no counter-intuitive ones. As we find from EDA, male customers are purchasing much more than female customers. The coefficient of total_purchase_log is positive which indicates there is a positive relationship between log total purchase and being male customers. On the contrary, we find from EDA that most of the customers from occupation_9 are female. The coefficient of occupation_9 is negative which implies there is a negative relationship between being male customers and from occupation_9, i.e., customers from occupation_9 are more likely to be female.

4.2 3.2 K-Nearest Neighbors

K-NN classification is implemented using the sklearn's KNeighborsClassifier Class. There is only one parameter that needs to be tuned:

- n_neighbors: number of neighbors to be used for kneighbors queries.

The test accuracy of KNN model is 0.72. From the confusion matrix, we can find that KNN is not performing as well as logistic regression. Both the two numbers on the diagonal indicating correct predictions are lower than the ones of logistic regression.

Figure 18. Confusion Matrix Using KNN

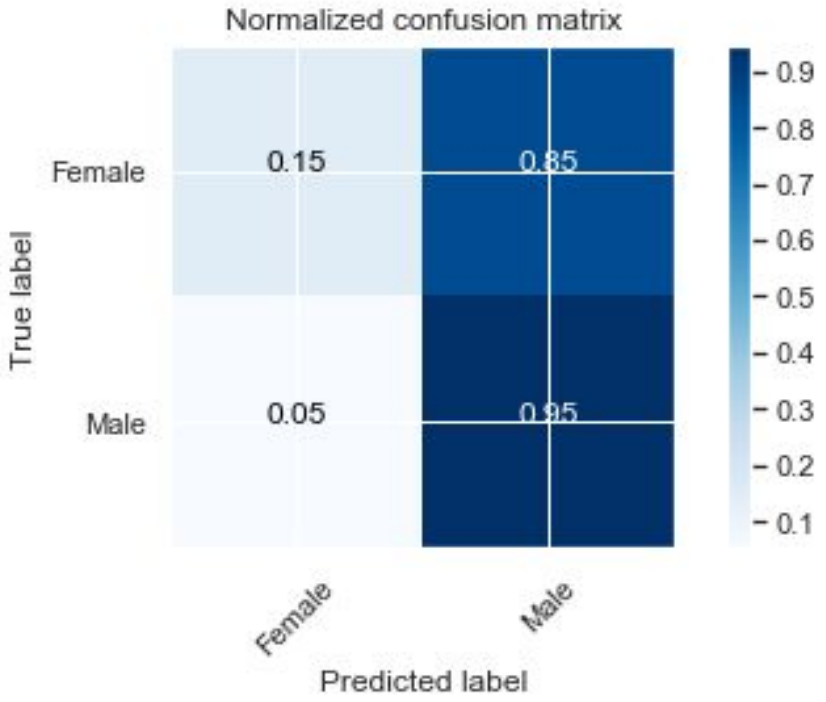
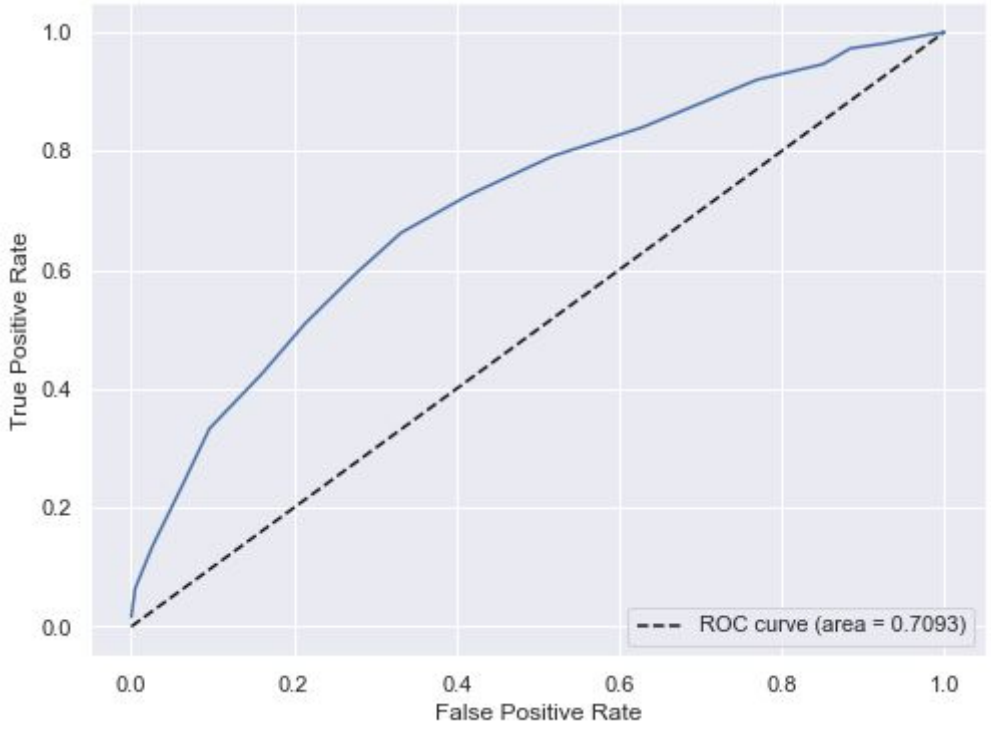


Figure 19. ROC Curve Using KNN



4.2.3.3 Support Vector Machines

First I implements a linear SVC using sklearn's LinearSVC class and tunes one parameter:

- C: penalty parameter of the error term

The test accuracy of linear SVC is 0.7651. The Area under ROC curve for this model is very close to the logistic regression, while logistic regression is a bit better.

Figure 20. Confusion Matrix Using Linear SVC

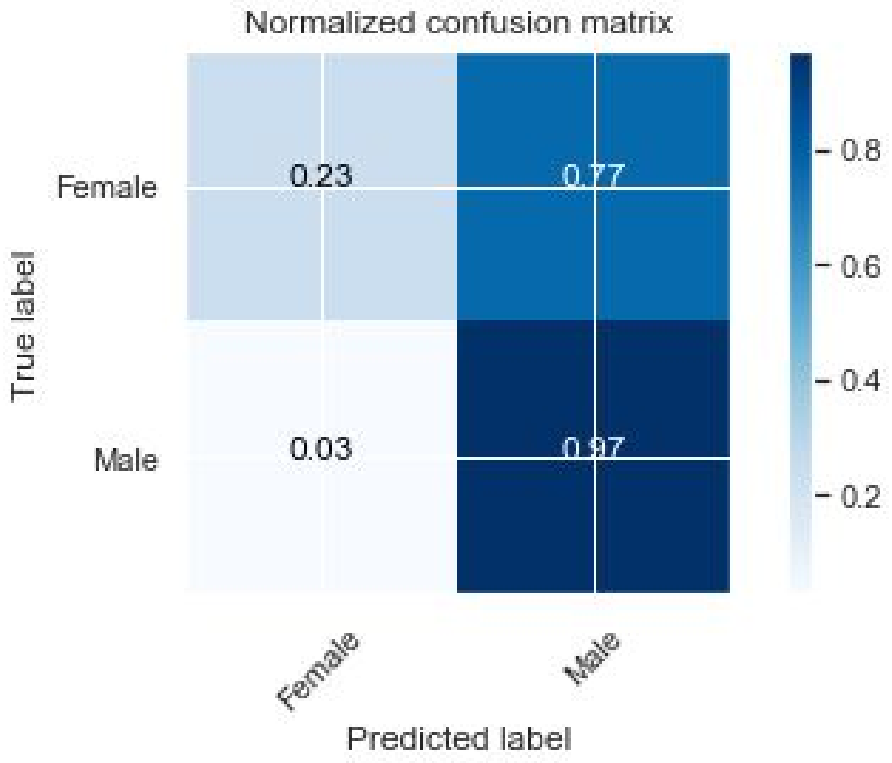
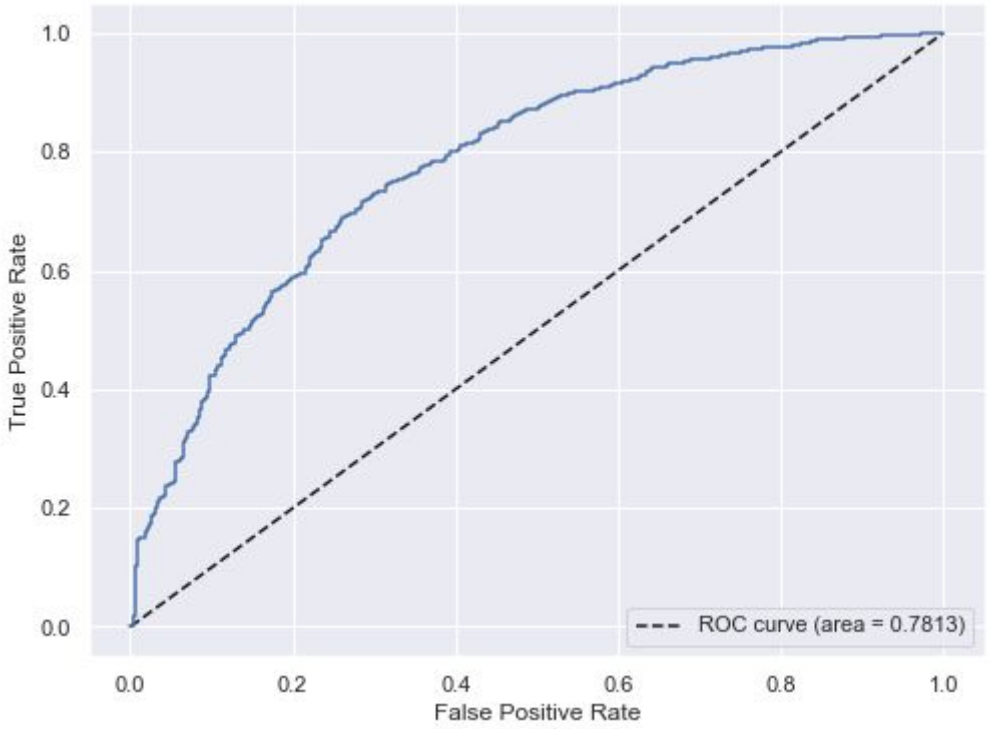


Figure 21. ROC Curve Using Linear SVC



Then I implements a SVC with kernel type of ‘rbf’ using sklearn’s SVC class. Two of the paramers are chosen to be tuned:

- C: penalty parameter of the error term
- gamma: kernel coefficient

The test accuracy is 0.7651 for SVC with kernal of ‘rbf’. The performance of the model is also further described in the following confusion matrix and ROC curve.

Figure 22. Confusion Matrix Using SVC with RBF Kernel

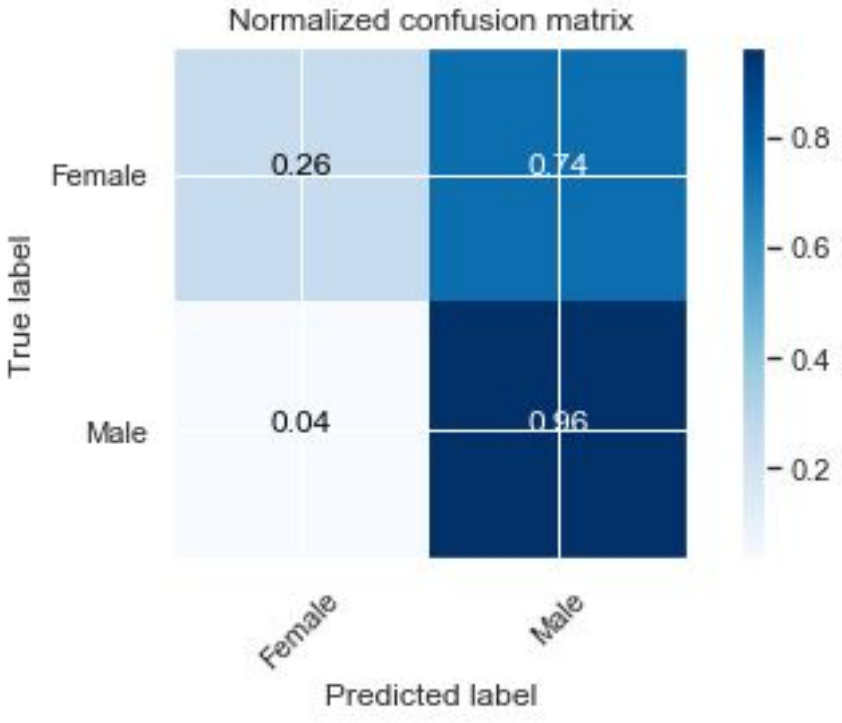
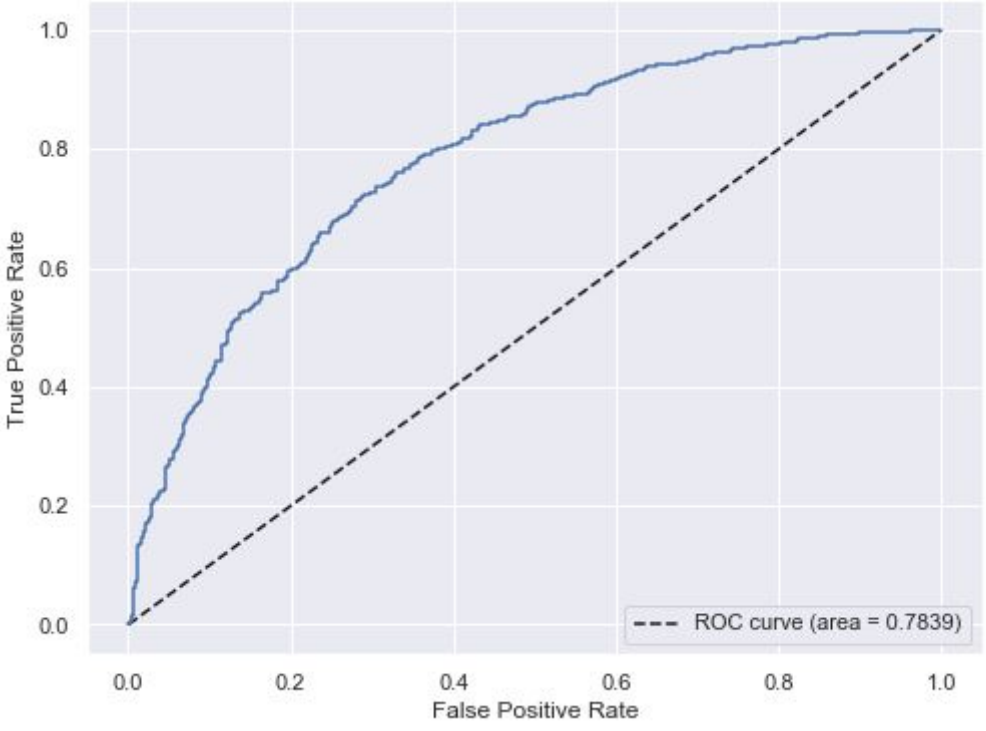


Figure 23. ROC Curve Using SVC with RBF Kernel



4.2.3.4 Random Forest

I used the sklearn's RandomForestClassifier class to implement Random Forest Classification. I chose n_estimators=200. Then another two parameters were chosen to be tuned:

- max_depth: the maximum depth of the tree.
- max_features: the number of features to consider when looking for the best split.

The test accuracy of Random Forest Classifier is 0.763. From the confusion matrix, we can see that true positive rate is quite high.

Figure 24. Confusion Matrix Using Random Forest

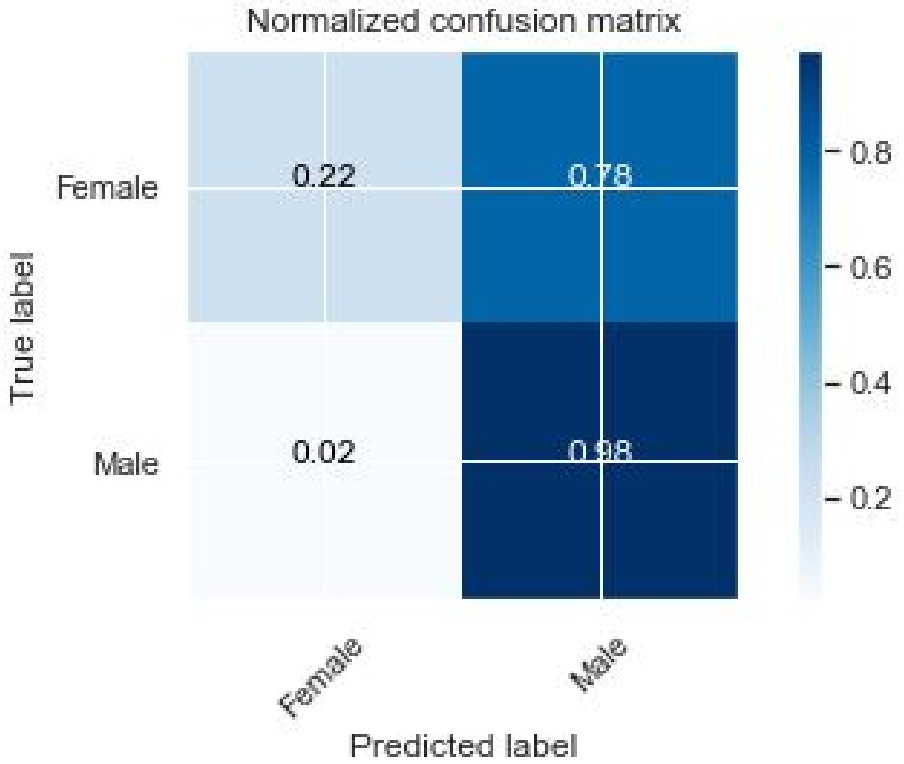
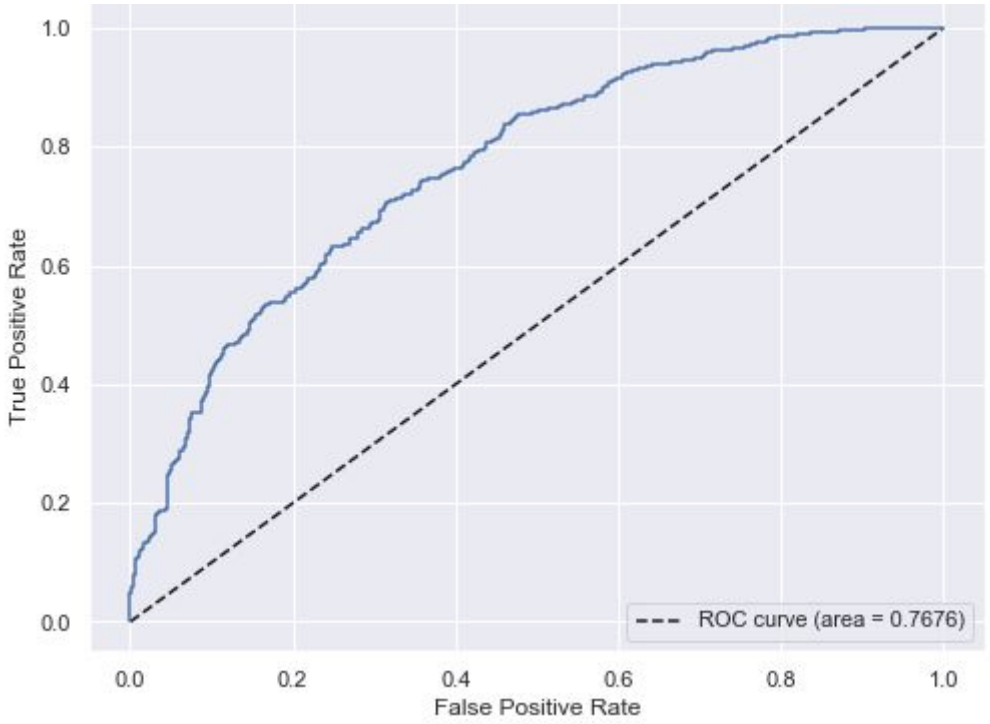
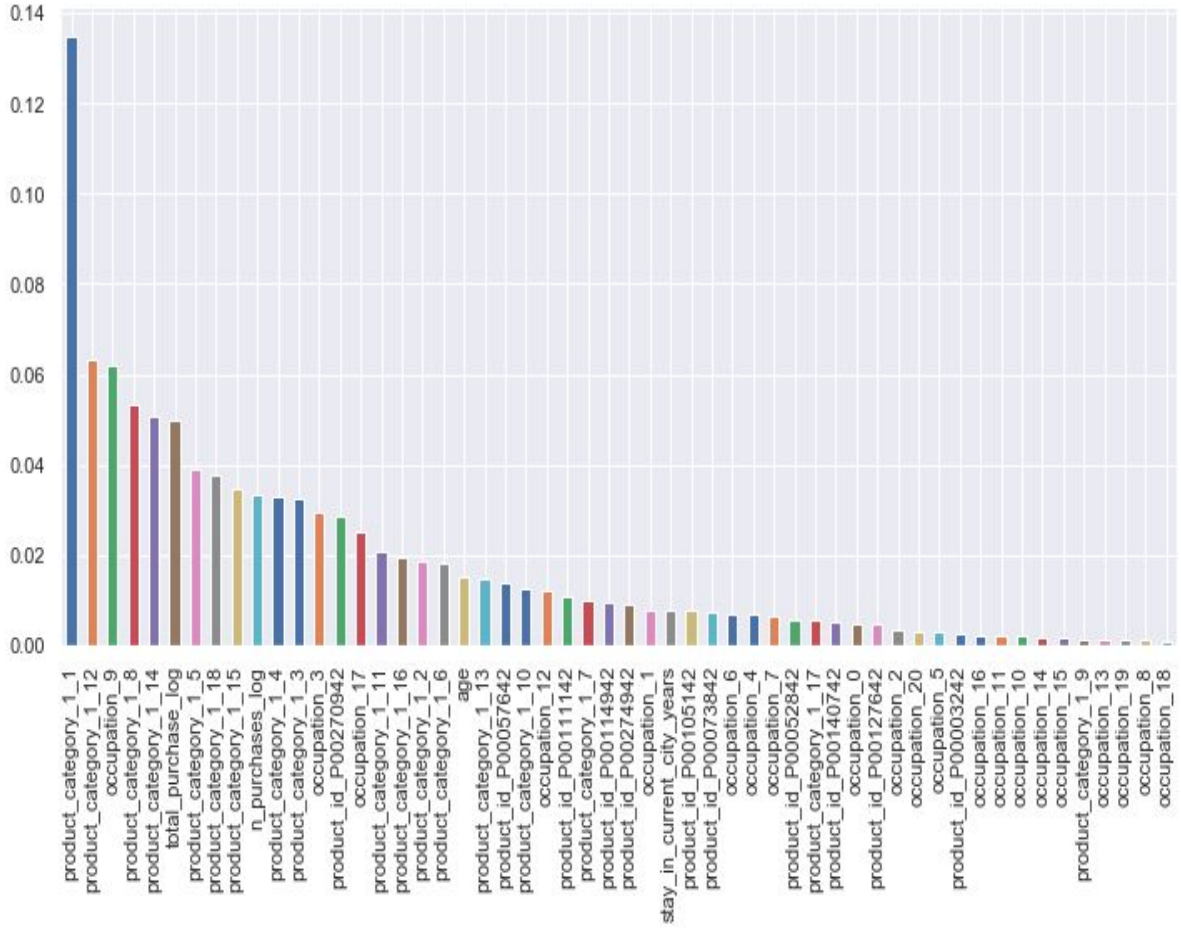


Figure 25. ROC Curve Using Random Forest



We can also use forests of trees to evaluate the importance of features on the classification task. The plot suggests that product_category_1_1 is the most informative feature in predicting the gender which makes sense because as can be seen from EDA, of all customers purchasing products in this category, only around 18% are female. In comparison, 28.3% of all customers are female. Thus we know male are more likely to purchase products in this category relatively speaking.

Figure 26. Feature Importances



4.2.3.5 Gradient Boosting Classifier

Last, I used the gradient boosting classification from sklearn's GradientBoostingClassifier class. I still chose n_estimators to be equal to 200. Then I chose to tune one parameter:

- max_depth: maximum depth of the tree.

The test Accuracy is 0.7725, almost as high as logistic regression. AUC under ROC is the highest, 0.7850.

Figure 27. Confusion Matrix Using Gradient Boosting Classifier

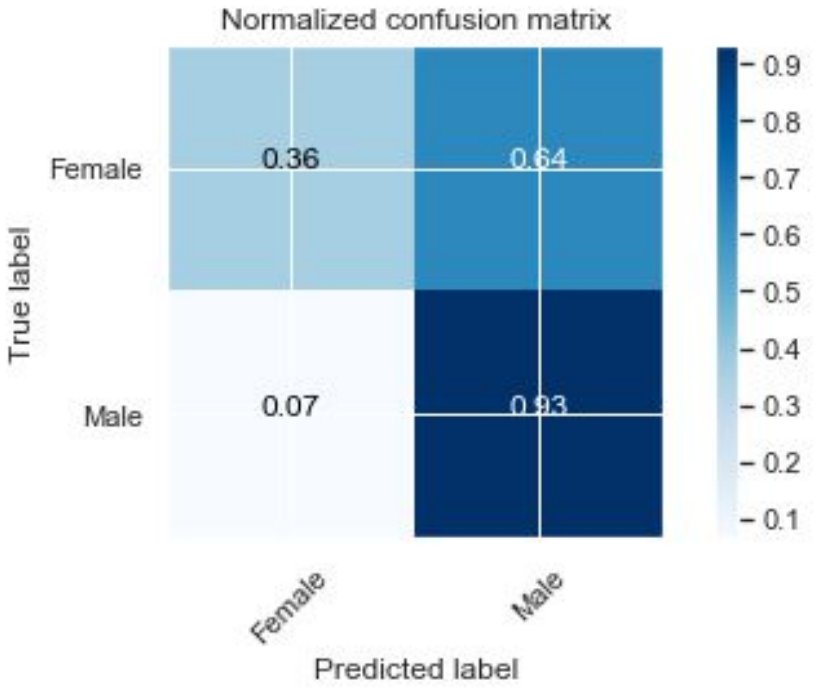
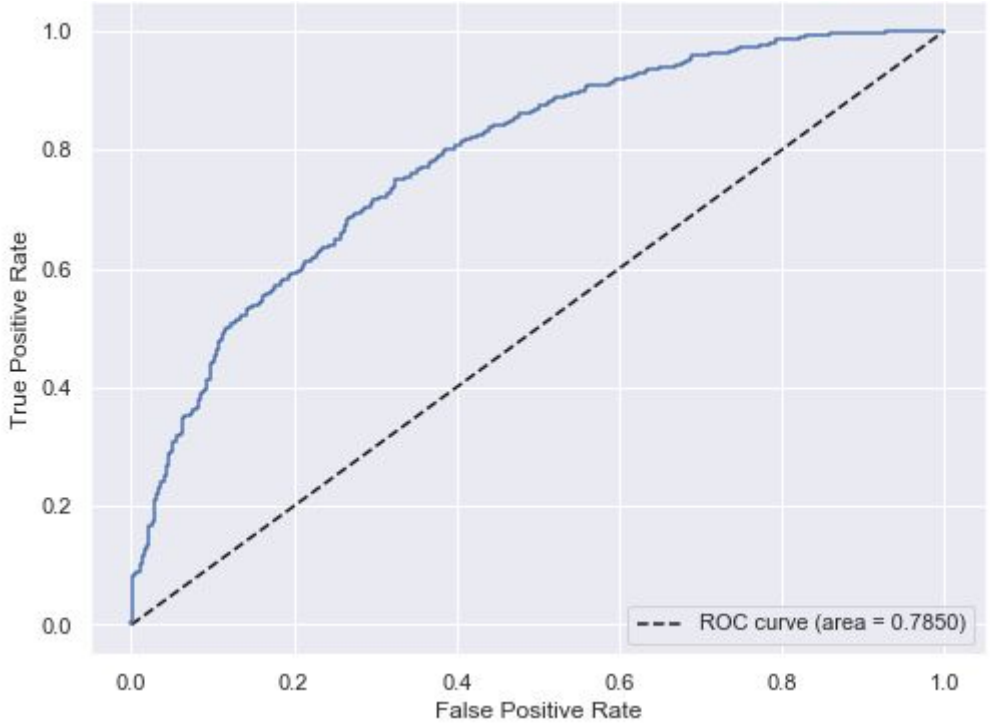


Figure 28. ROC Curve Using Gradient Boosting Classifier



4.2.4 Model Comparison

I have presented the accuracy score, confusion matrix, and ROC curve for every classification model used in the project. Here I will add their average precision, recall and f1 score to the summary table below.

Table 3. Summary of Model Performance

	Logistic Regression	KNN	Linear SVC	SVC with 'rbf' Kernel	Random Forest	Gradient Boosting
Recall	0.77	0.72	0.77	0.77	0.76	0.77
Precision	0.77	0.68	0.77	0.76	0.77	0.76
F1 Score	0.74	0.66	0.72	0.72	0.71	0.75
Accuracy	0.7739	0.7230	0.7651	0.7651	0.7630	0.7726
AUC	0.7821	07093	0.7813	0.7839	0.7676	0.7850

5. Conclusion

In this project, I use supervised learning to classify customer gender. Our data set provides useful features such as customers' occupation, living area, age, purchase records including products id, category, and purchase amount. Some feature engineering and selection work was done to prepare the data set used to train various classification models.

Generally, classification models are able to label female and male customers with an accuracy of more than 70%. Different models performs different, some with higher ability to identify female customers, while some with higher ability to identify male customers. The best two performing models are logistic regression and gradient boosting classifier. I use the default threshold 0.5 to do the classification

and get a better predictive result of male customers. Male customers are the main purchase power of this store, thus it is reasonable that we focus more on labelling male customers correctly and target them better to attract and retain them.