Predicting the Gender of Customer

Caixuan Sun

Springboard Capstone Project 1
January 2019

What does Retail care?

- Win and retain **customers**
- Improve business efficiency

How can Data Science Help?

- Understand who you are selling to?
- Personalized recommendations
- Personalized offers
- Targeted campaigns
- Uncover trends and cross-selling opportunities

The Task:

- To predict the gender of customers to help retail stores improve performance.

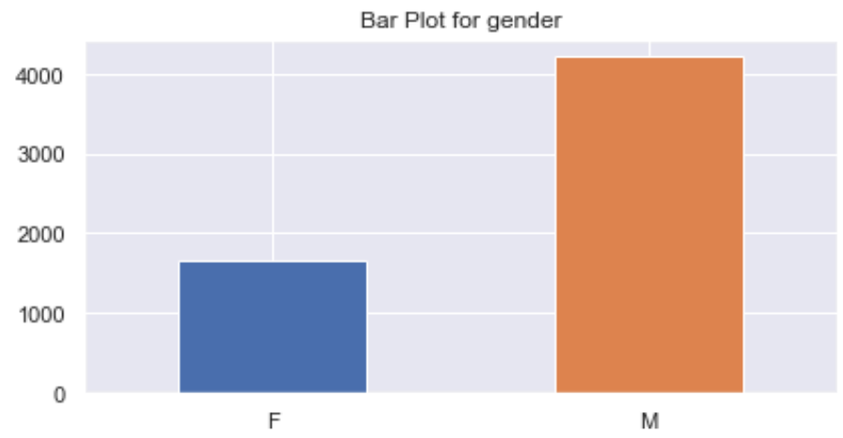
Data Description

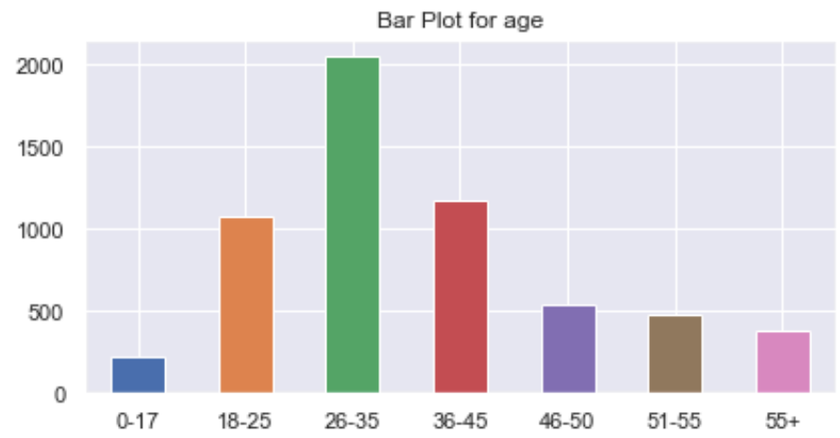
- Data acquired from Kaggle
- Transactions data from a retail store for a period of one month
- Number of observations: **537,577**
    - Each observation is a purchase record of one particular product by one customer
- Number of features: **12**
    - <u>Customer level features:</u>
        - Gender, age, occupation, city_category, marital status, stay in current city years
    - <u>Product level features:</u>
        - Product id, product category, purchase amount

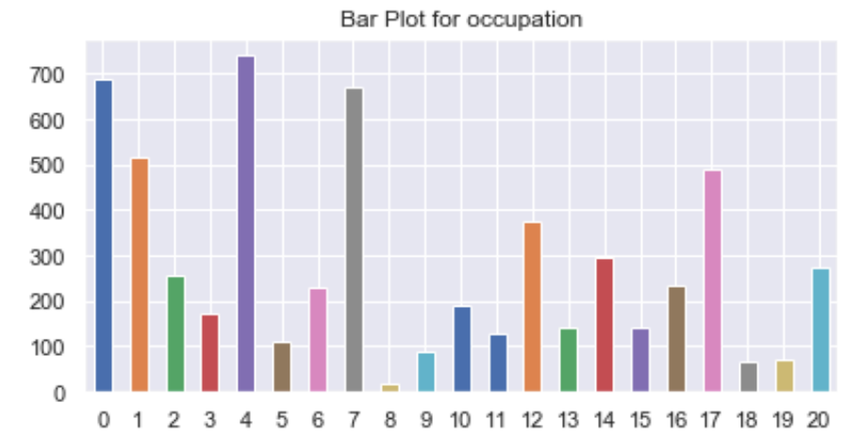
Data Cleaning and Wrangling

- How many customers? 5,891
- How many products? 3,623
- Missing values were identified and filled
- Data is aggregated into customer-level

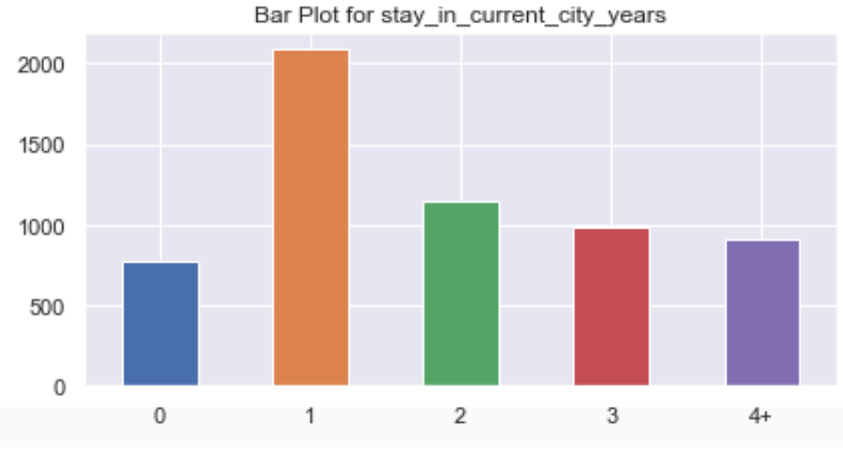
EDA

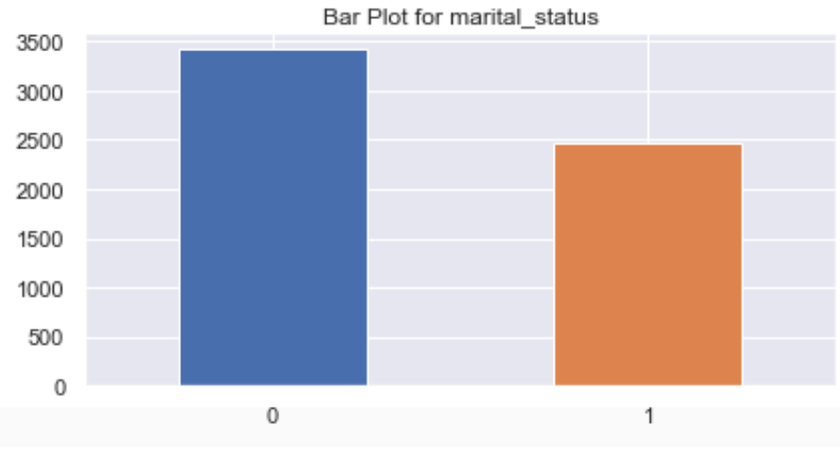

Bar Plot for gender
4000
3000
2000
1000
0
F
M


Bar Plot for age
2000
1500
1000
500
0
0-17
18-25
26-35
36-45
46-50
51-55
55+

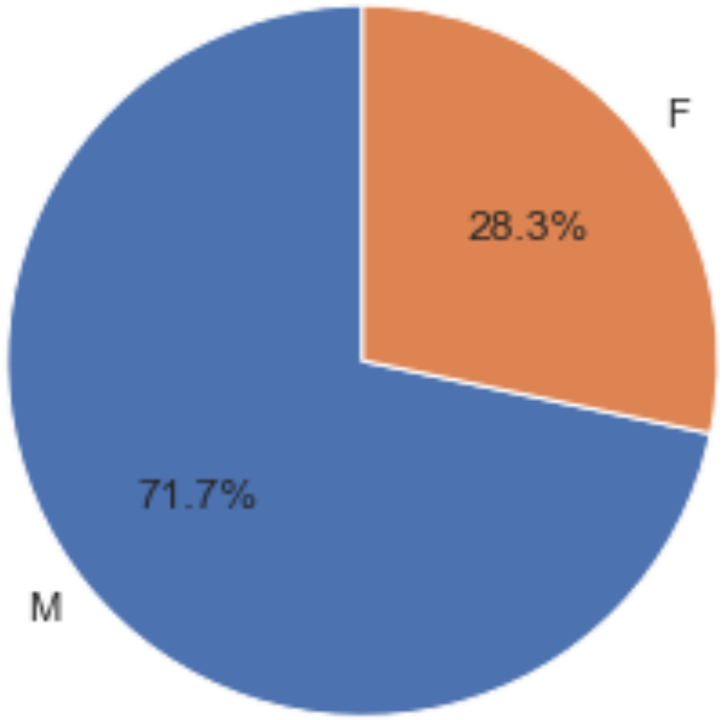

Bar Plot for occupation
700
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


Bar Plot for city_category
3000
2500
2000
1500
1000
500
0
A
B
C


Bar Plot for stay_in_current_city_years
2000
1500
1000
500
0
0
1
2
3
4+

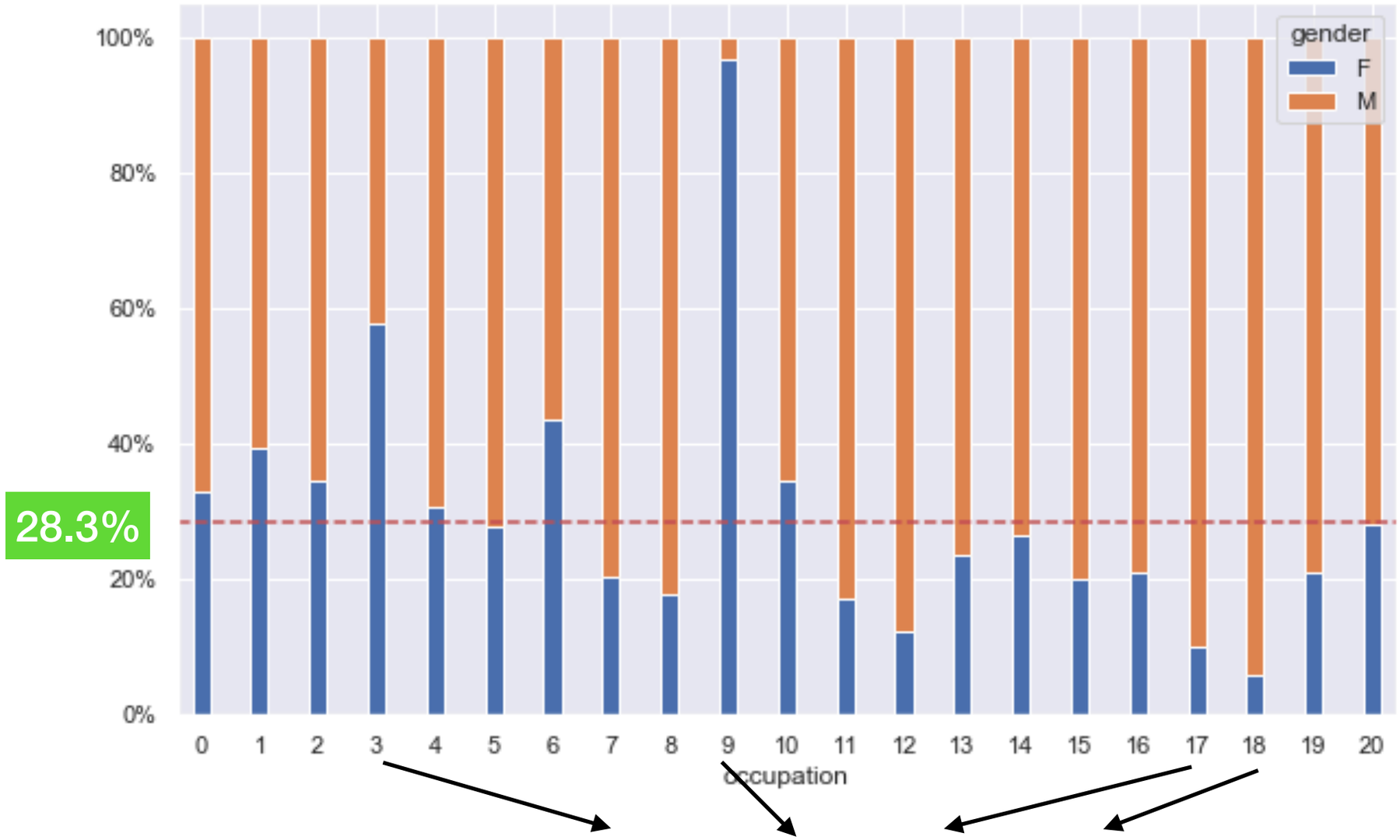

Bar Plot for marital_status
3500
3000
2500
2000
1500
1000
500
0
0
1

EDA: Target Variable ‘Gender’



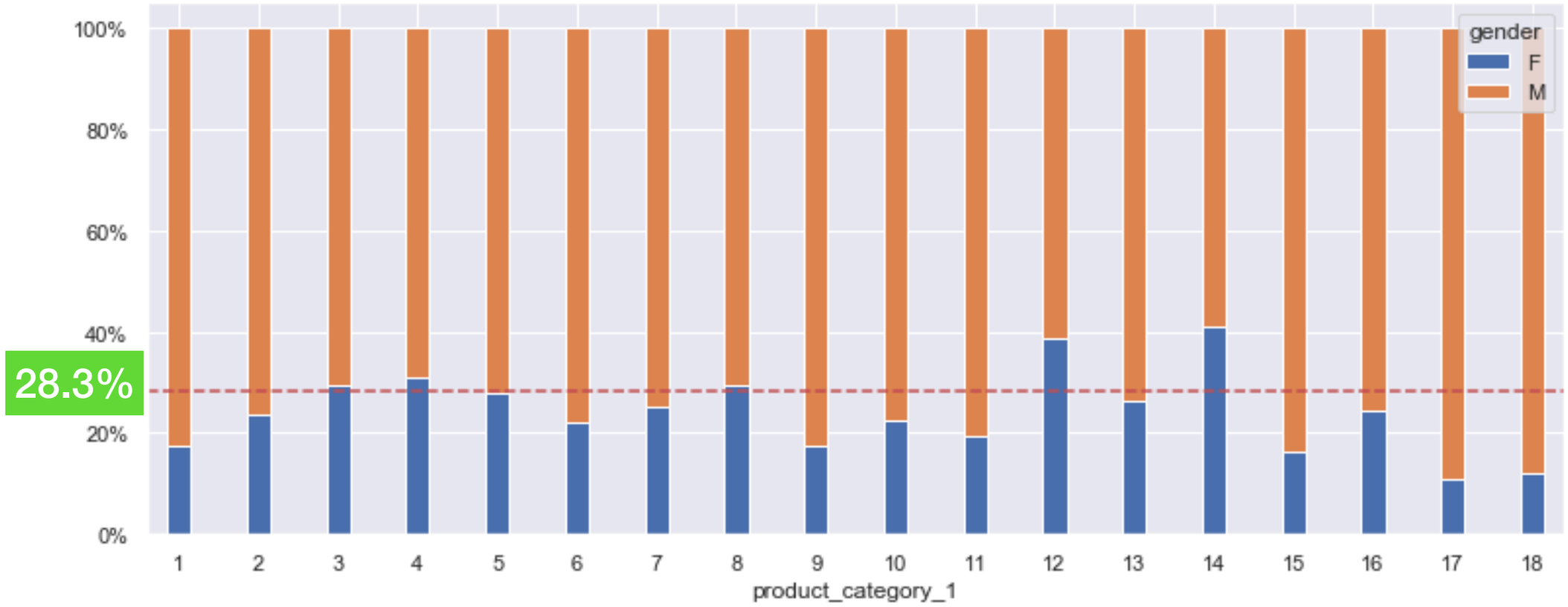
	Count	Gender Proportion	Purchase Proportion
Female	1666	28.3%	23.2%
Male	4225	71.7%	76.8%

Gender and Occupation



These four categories may provide useful information for the prediction

Gender and Product Category



Machine Learning Modeling

- Supervised learning
- Binary Classification: 1 for Male and 0 for Female

- Models Used:
  - Logistic Regression
  - KNN
  - SVM
  - Random Forest
  - Gradient Boosting

- Evaluation Metrics Used:
  - Accuracy Score
  - Confusion Matrix
  - Classification Report
  - ROC Curve
  - Area Under ROC Curve

Modeling Steps

1. Data Preprocessing:
   - Label encoding for categorical variables
   - Scaling numerical variables
   - Split data into training and test sets

2. GridSearchCV for hyperparameter tuning:
   - 5 fold cv
   - Using training set

3. Train Classifier using optimal parameters on the training set

4. Performance evaluation using holdout set

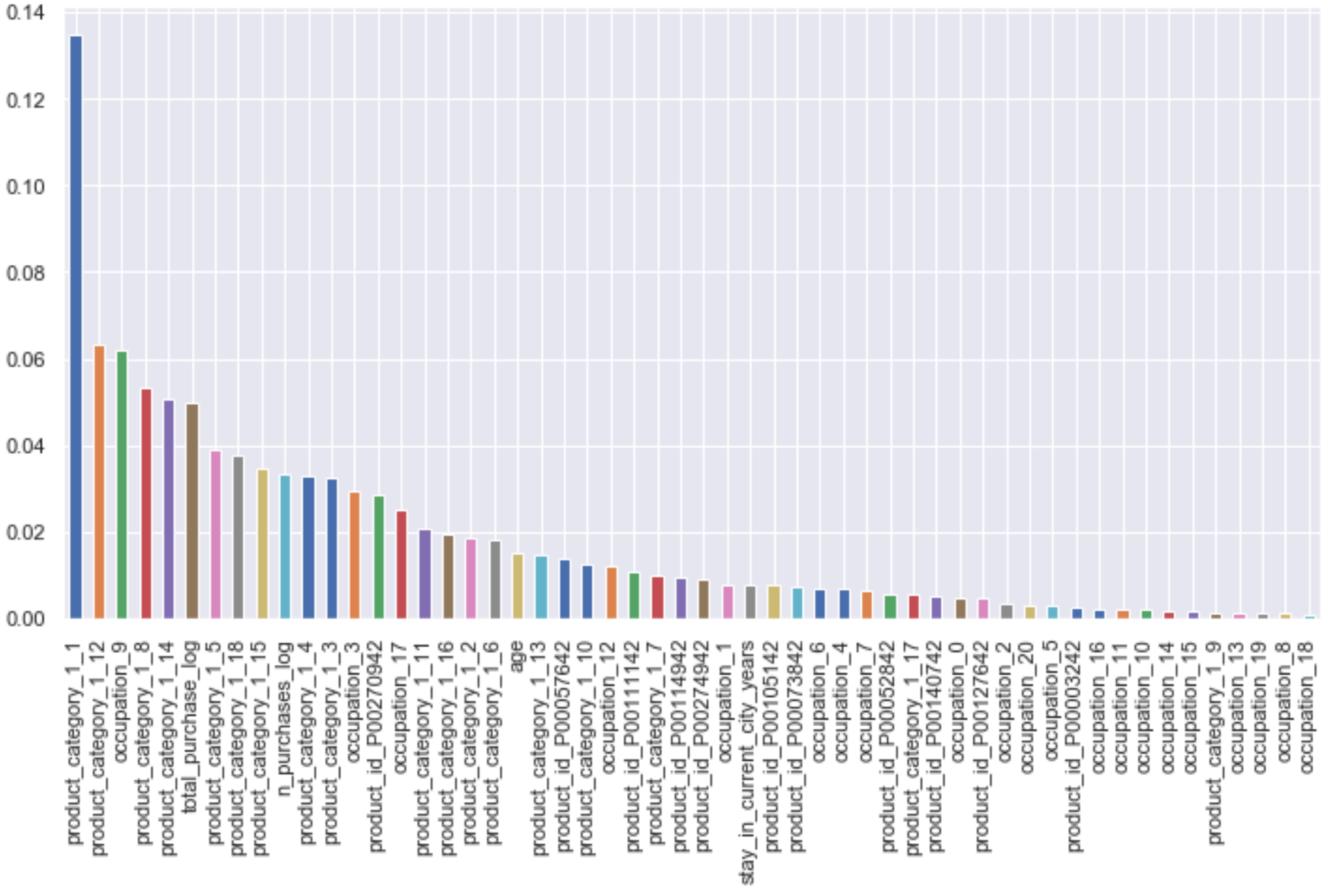
Model Performance Comparison

	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	0.7093	0.7813	0.7839	0.7676	0.7850

Worst Performance Model: KNN

Best performance model:
Logistic regression
&
Gradient Boosting

Feature Importances



Conclusions

- Out of 5 supervised classification models, Gradient Boosting Classifier provided the best prediction based on ROC AUC.

- Limitations:
  - Two features product_category_2 and product_category_3 are not included in the feature set, which may provide useful information.

  - All models are not performing well at labelling female customers which may because the imbalance sample between female and male customers.

Thank you!

Caixuan Sun
Email: suncaixuan@gmail.com
Project Repository at Github: https://github.com/caixuansun/Springboard/tree/master/capstone_project_1