Customer Segmentation Classification

Data Mining - Fall 2022

Claraestela Torres, Bristow Richards, Aditya Singh

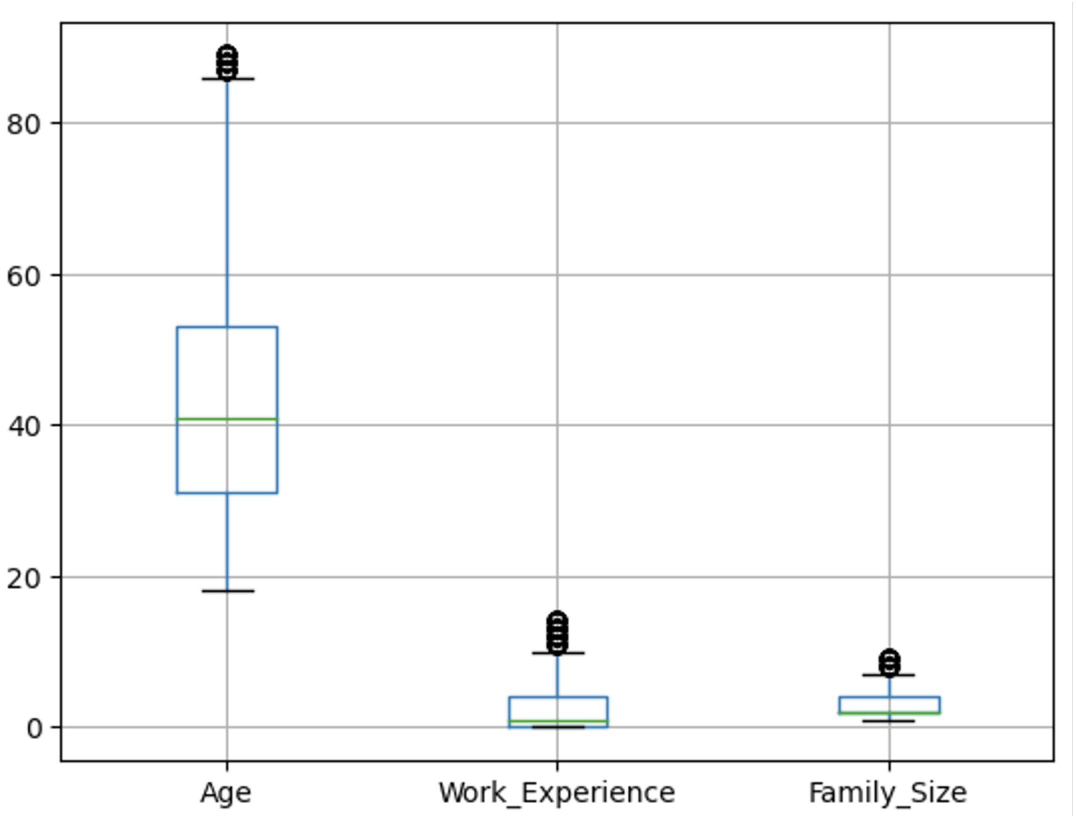
Data Overview

- “Customer Segmentation Classification” uploaded to Kaggle by user Kash
  - Link

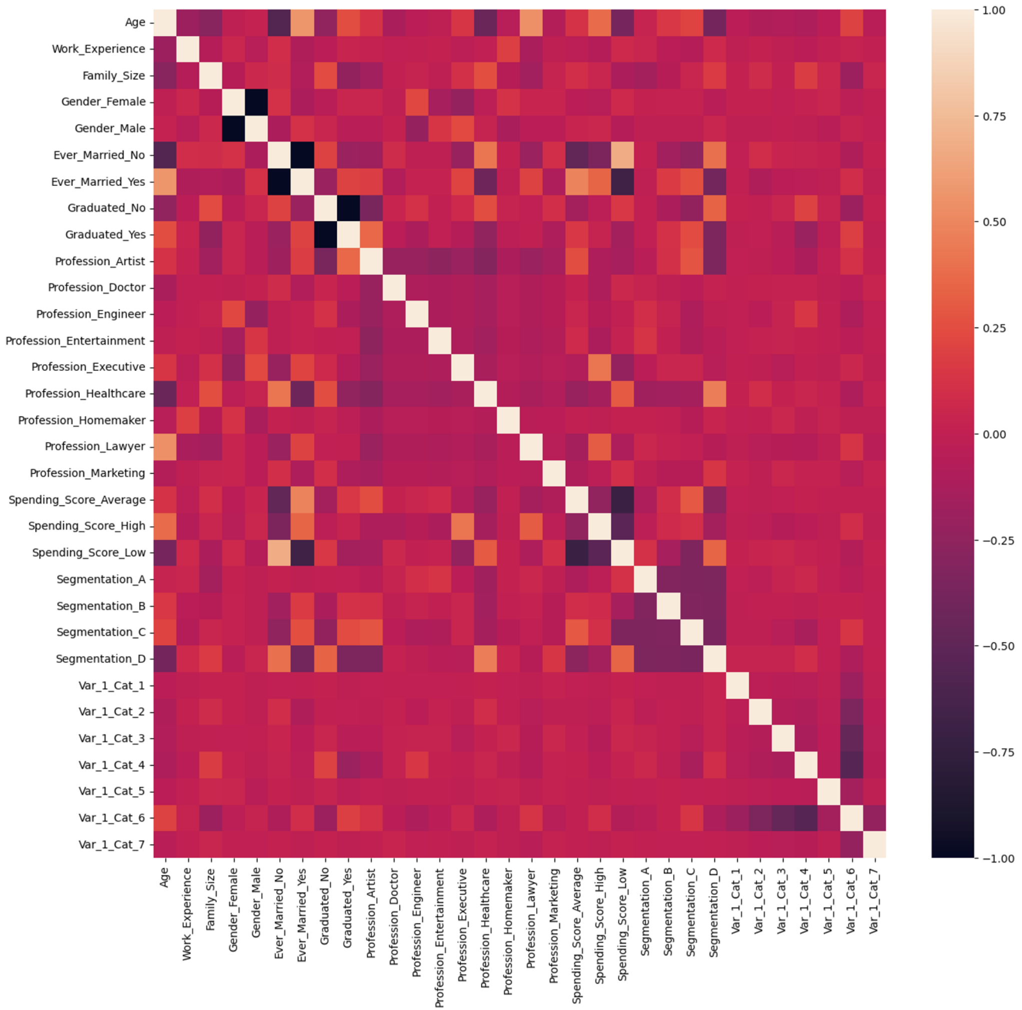
- This data had 11 columns: 1 id column, 1 target variable, and 9 predictors
  - **ID** - Unique ID
  - **Gender** - Gender of the customer
  - **Ever_Married** - Marital status of the customer
  - **Age** - Age of the customer
  - **Graduated** - Is the customer a graduate?
  - **Profession** - Profession of the customer
  - **Work_Experience** - Work Experience in years
  - **Spending_Score** - (target 2) Spending score of the customer
  - **Family_Size** - Number of family members for the customer (including the customer)
  - **Var_1** - Anonymised Category for the customer in the training data
  - **Segmentation** - (target 1) Customer Segment of the customer

- The data had around 10,500 observations, split roughly 80%-20% into training and testing. Roughly 20% of the data had null values.

Continuous Data



Correlation Matrix



Task 1: Customer “Segmentation” (Classification)

KNN Classifier - 2 neighbors

	precision	recall	f1-score	support
D	0.340	0.403	0.369	692
B	0.198	0.260	0.225	450
C	0.241	0.270	0.255	381
A	0.497	0.249	0.332	631
accuracy			$\boxed{0.305}$	2154

Decision Tree - Leave nodes: 20

	precision	recall	f1-score	support
D	0.35	0.37	0.36	692
B	0.29	0.25	0.27	450
C	0.26	0.34	0.30	381
A	0.44	0.38	0.41	631
accuracy			$\boxed{0.34}$	2154

Naive Bayes

	precision	recall	f1-score	support
D	0.327	0.247	0.281	692
B	0.236	0.076	0.114	450
C	0.230	0.520	0.319	381
A	0.405	0.403	0.404	631
accuracy			$\boxed{0.305}$	2154

Random Forest

	precision	recall	f1-score	support
D	0.32	0.30	0.31	692
B	0.22	0.07	0.10	450
C	0.25	0.45	0.32	381
A	0.44	0.46	0.45	631
accuracy			$\boxed{0.33}$	2154

Task 2: Customer Spending Score (Classification)

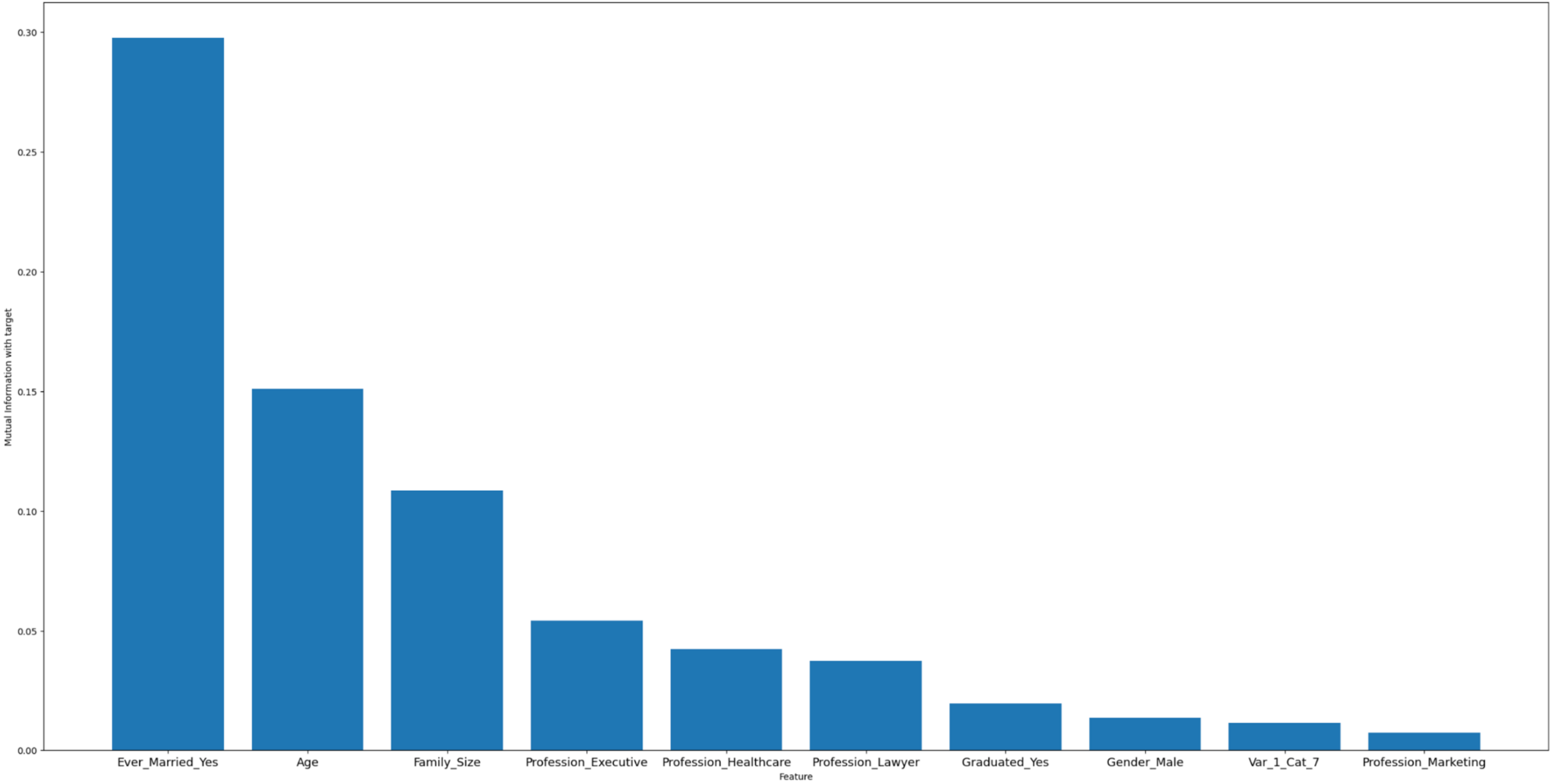
Setup - Multi-class classification

Target: Spending Score -> Low, Mid, High

Features: All features except segmentation

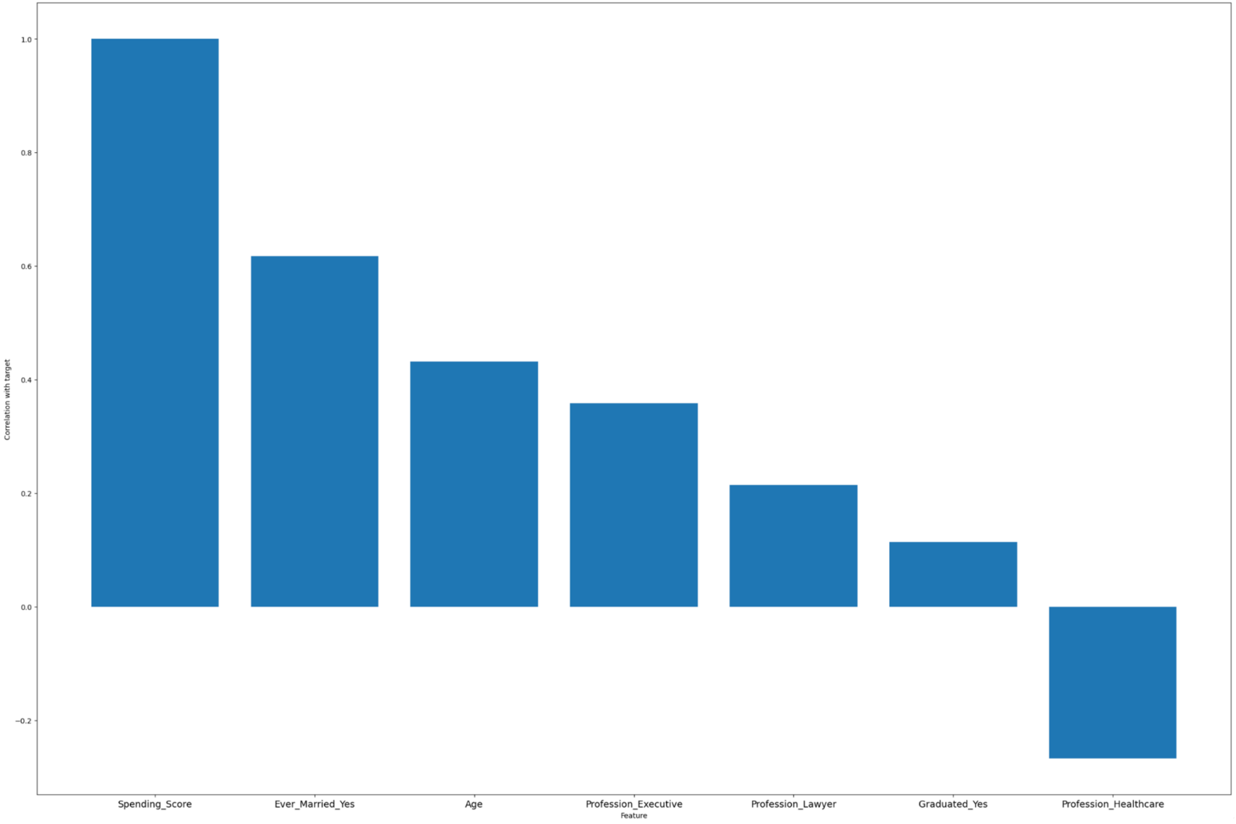
Let’s look at important features before modelling

Information Gain

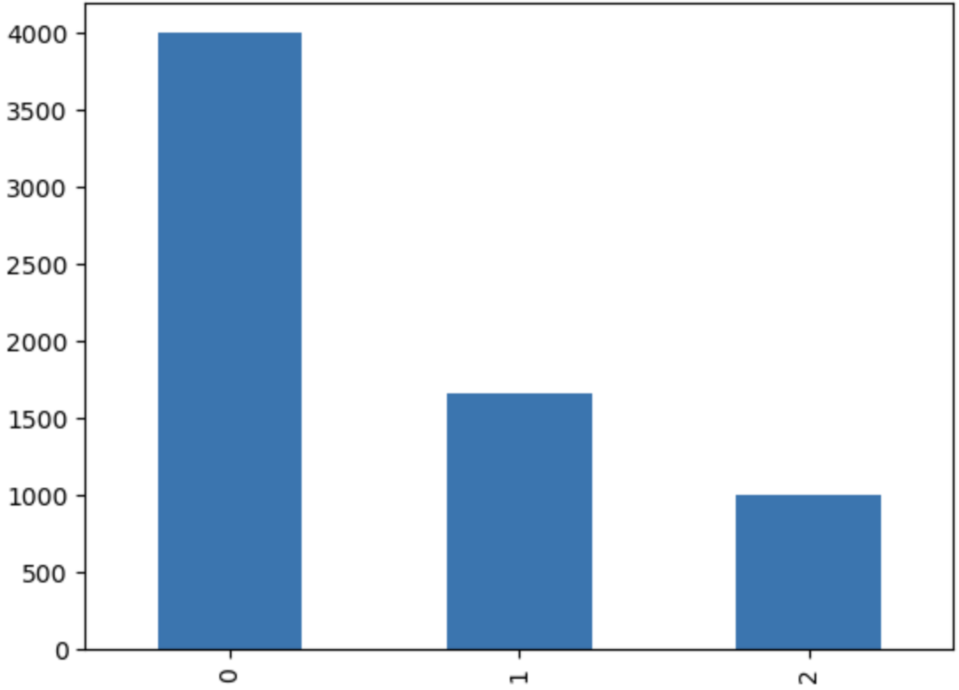


Correlation with Target

For Features with abs(correlation)>0.1



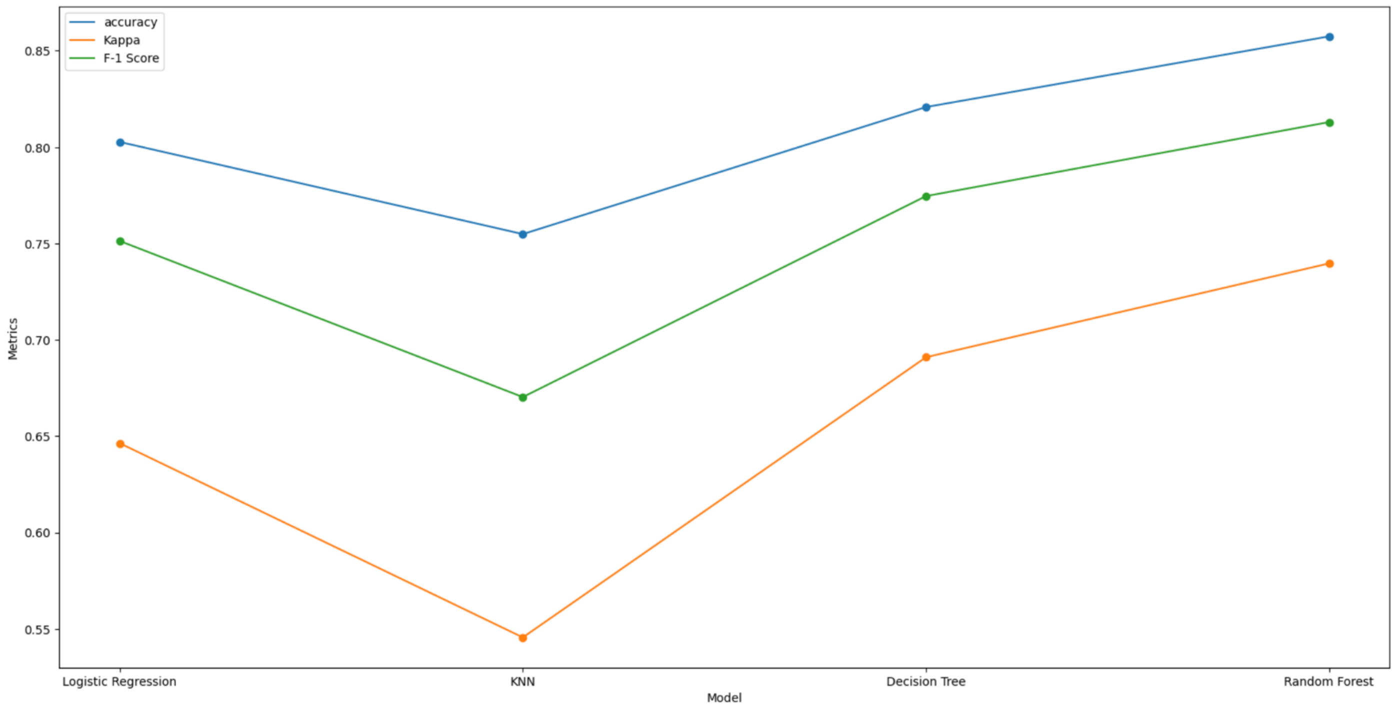
Label Imbalance



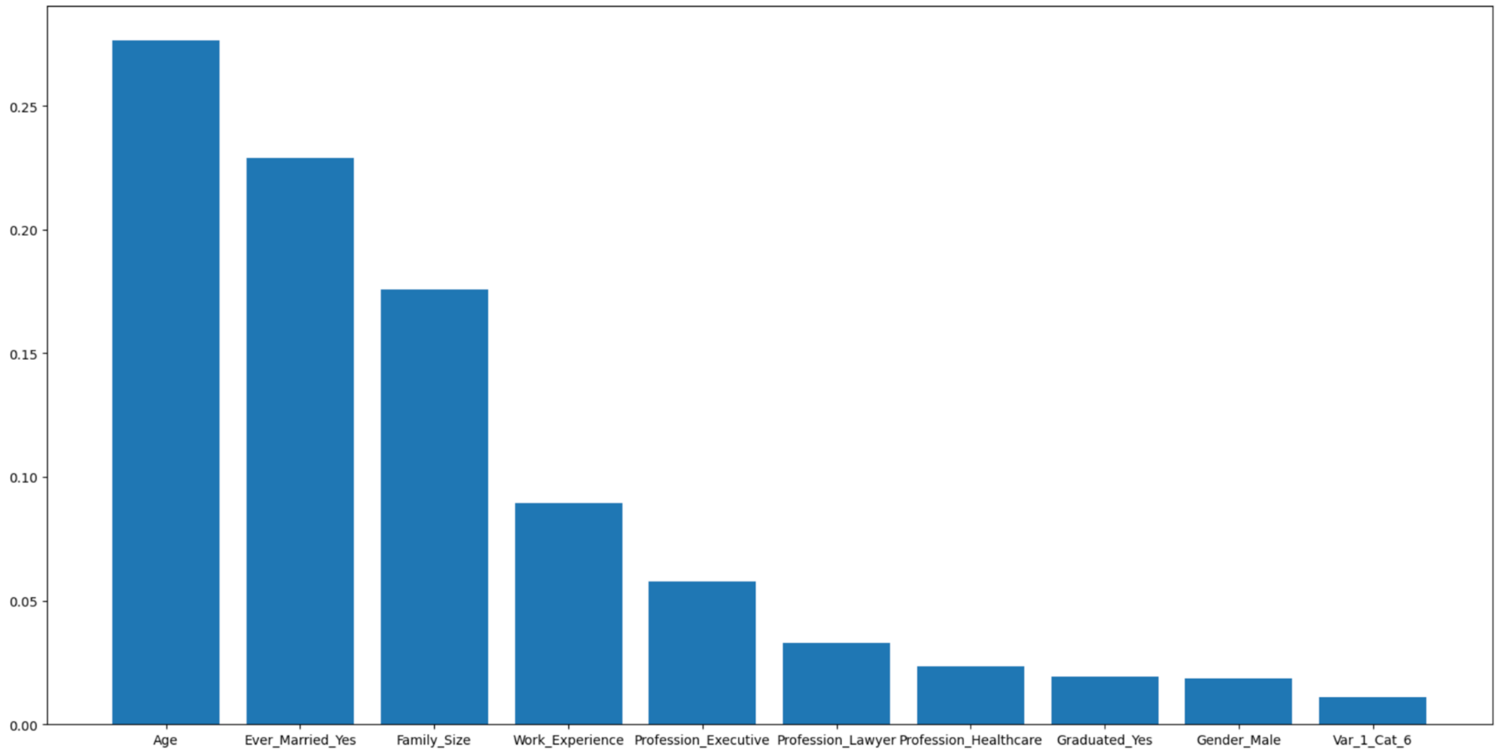
0: Low spenders,
1: mid spenders
2: high spenders

We see that there are many more low spenders

Comparing the classification models

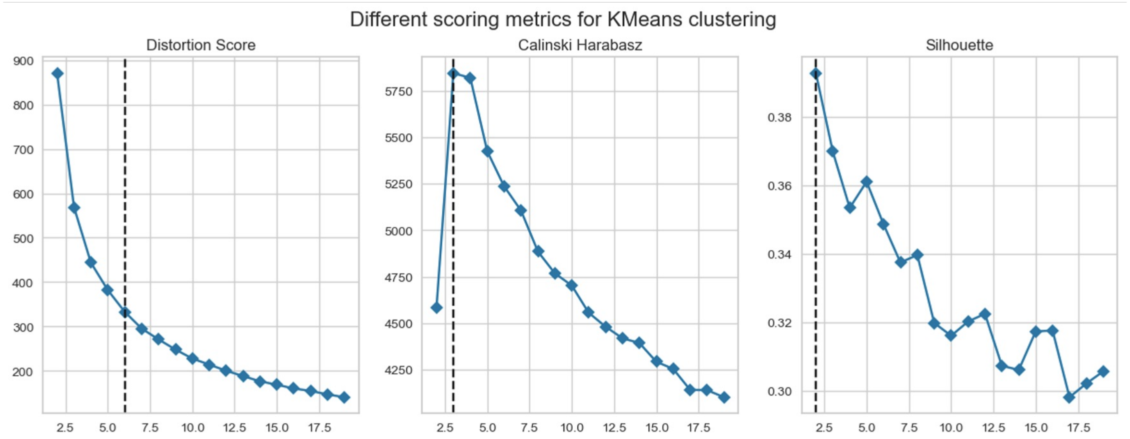


Random Forest Feature Importance



Task 3: Exploratory Clustering

Clustering attempts



- Scoring metrics for clustering could not converge on ideal value for **k**
- Normalizing data or ignoring categorical data did not help
- It was not possible to explore different clusters for qualities the business would be interested in

Closing Remarks

Closing Remarks

- Too many categorical variables made our data sparse
- Clustering algorithms we learned perform better on continuous data
- Classification models struggled on classifying “Segmentation”
  - Maybe this is because the segmentation is arbitrary and fundamentally not correlated strongly with features -> business problem
- Classification models succeeded on classifying “Spending_Score”