Supervised Learning - Customer Segmentation

Mohsin Mohammed (100744768)

Faculty of Science, Engineering & Information Technology

Introduction to Machine Learning (COSC-21000)

Dr. Ruba Al Omari

December 11, 2022

Introduction

The research problem includes an automobile company that is trying to classify new prospective customers into four categories namely A, B, C, and D. The company has already classified its existing customers with 8068 unique IDs into these four categories and wishes to do the same for the prospective customers. Our job is to use the current customer base and their classification to build and train different machine learning algorithms and predict the target class for the new prospective customers.

Accurately segmenting customers based on common characteristics allows a company to focus on creating products that are more relevant to and serve the needs of each group. It allows a company to focus on groups of similar customers rather than on individual customers. It enables a company to allocate its marketing and advertising resources efficiently. In the long run, businesses can better serve their customers and increase their brand value and revenues.

Related Works

Related work 1

The author begins the machine learning process with data preprocessing. The first step involved using principal component analysis in removing columns that did not add any value in building the machine learning algorithm. The algorithm uses singular value decomposition to perform an orthogonal transformation to create components that are highly correlated (Rosas, 2019). The author then proceeds to cluster the unlabeled data using the Kmeans algorithm. This helped the author segment the customers into 15 clusters as indicated by the elbow method. The segments with large distribution were analyzed further. The author then proceeds to train the labeled data with classification algorithms.

Related work 2

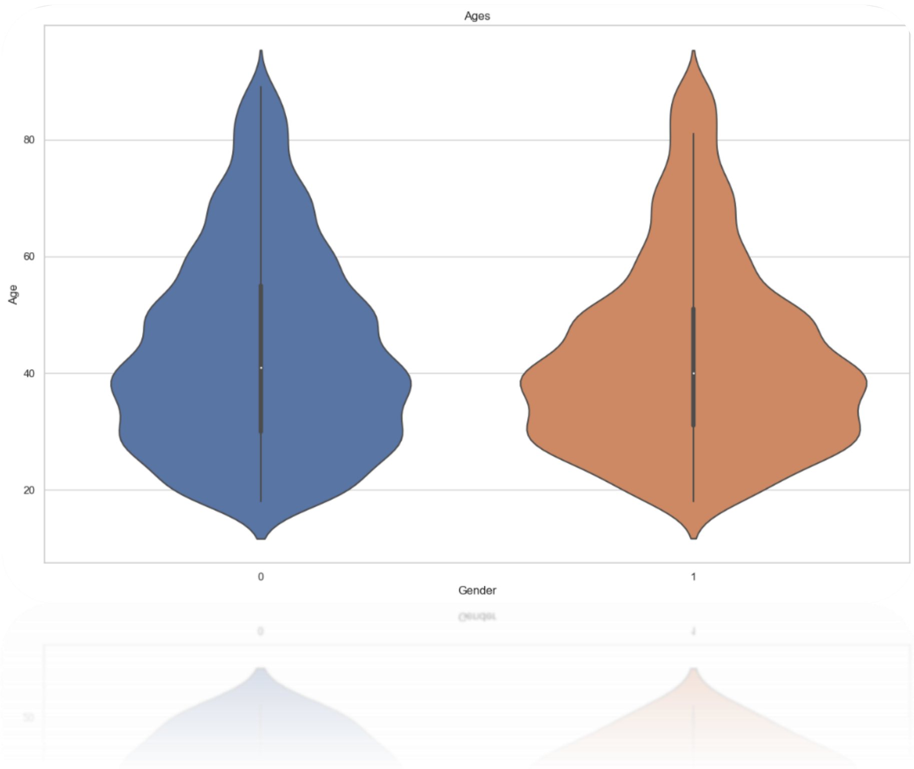
The author demonstrates how to use k means clustering which is an unsupervised machine learning algorithm to classify customers on unlabelled data set. The algorithm assigns k random centroids randomly and each data point is assigned to the closest centroid. The mean for all the data points is calculated and the centroids are adjusted to minimize the sum squared error. The algorithm does this iteratively until convergence or until it reaches the maximum iterations limit (Arvai, n.d.). The data was preprocessed using pandas and NumPy, following which, the Kmeans algorithm made available by the scikit learn library was used to fit the data. The optimal number of k which denotes the number of clusters was determined using the elbow method. The data is then classified using the centroids that minimize the sum squared error.

Related work 3

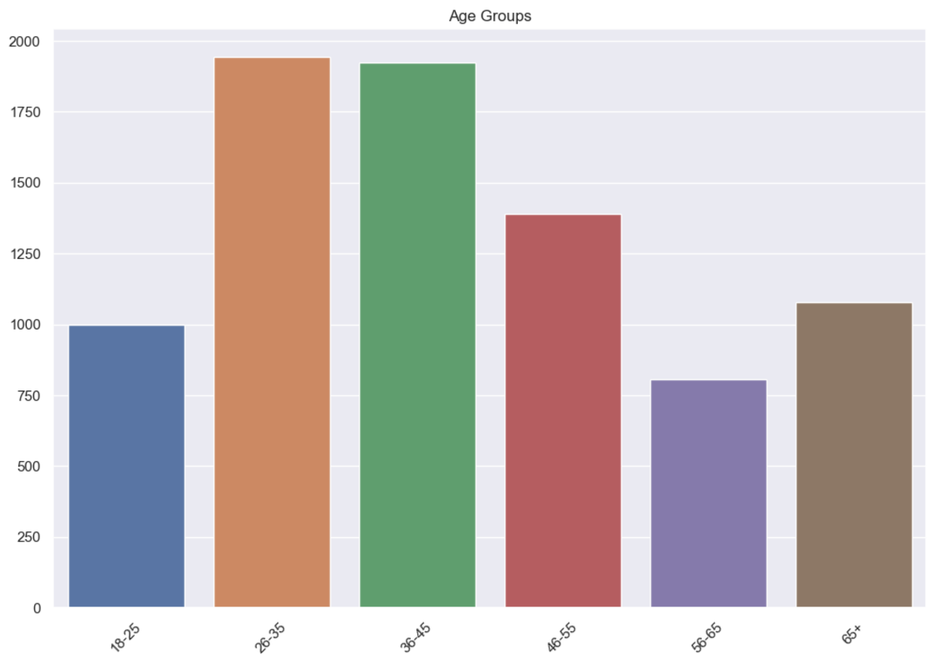
The author uses exploratory data analysis to learn how the data is distributed by plotting key findings. The next steps involve cleaning the data by imputing the missing values with the mean of the numerical attributes and the mode for categorical attributes. Since the data is unlabeled, the author uses Kmeans clustering to create clusters that represent customer segmentation. The elbow method was used to determine the optimal value for k which denotes the number of clusters. Each cluster was further analyzed using exploratory data analysis to better understand the distribution and make meaningful connections.

Exploratory Data Analysis

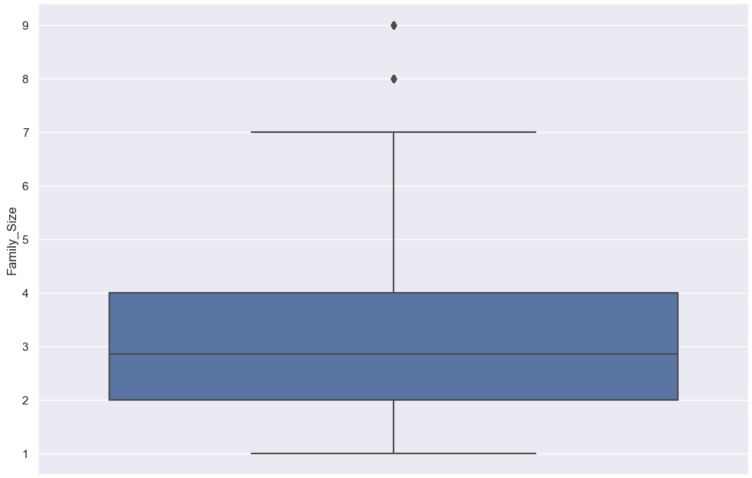
The dataset being used is a synthetic dataset that was used for a competition by a platform named Analytics Vidhya. The link can be found here. The dataset came to split into train and test sets. The training set is labeled while the testing set is unlabeled. The data set consists of customers with unique IDs. The competition required predicting the labels for the test set. Exploratory data analysis was done on the train set which consisted of 8068 instances. The describe() method was used to learn that 75% of the customers were below 53 years of age and 25% were under 30 years of age. A violin plot on gender and age revealed that the company had slightly more female customers than males. 0 represents the male distribution while 1 represents the female distribution.



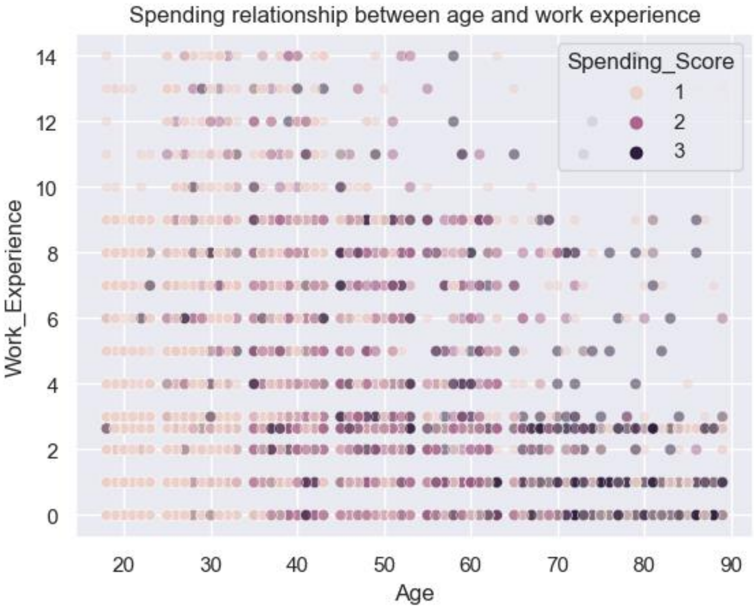
It was also found that the youngest customer is 18 and the oldest customer is 89 years old with the distribution being the highest in the age groups between 26-45 years of age. By visualizing the different age groups using a bar plot, the company can tailor products and services to cater to the needs of specific groups which ultimately leads to customer satisfaction, customer retention, and increased revenues.



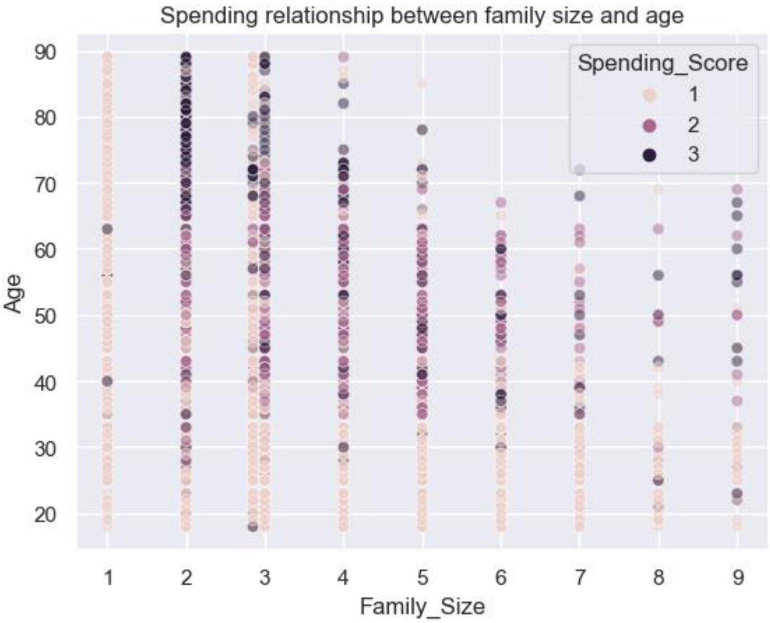
A box plot was also used to understand that 75% of the customers have 4 people per household and an average household family consists of 3 people. 25% of the customers had 2 people per household. Some customers have family households of 8 and 9 which fall outside the normal distribution and are considered outliers. EDA helps understand where the outliers lie because some algorithms are sensitive to outliers and are prone to overfit the data.



A scatter plot of 'Age' and 'Work Experience' against the 'Spending Score' attribute reveals that a customer's spending score improves as they get older. Specifically, customers between 40 to 80 years old and have up to 4 years of work experience are big spenders. We can see that customers up to approximately 30 years of age have lower spending scores and the years of work experience do not affect this demographic. We can deduce from this graph that the company's customers under the age of 30 regardless of their years of work experience are not big spenders. The spending score can be interpreted as 1 is low, 2 is average and 3 is high.



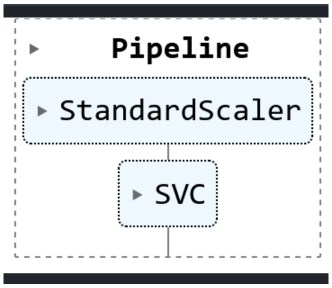
A scatter plot of 'Family size' and 'Age' against the 'Spending Score' attribute reveals that a customer's family size affects their spending behaviors, customers that have a family size of 2 and between the age group of 65 to 89, spend a lot of money and have a higher spending score. As the family size increases the spending scores are lowest up to the age of 30. The spending scores are average between the ages of 40 and 60.



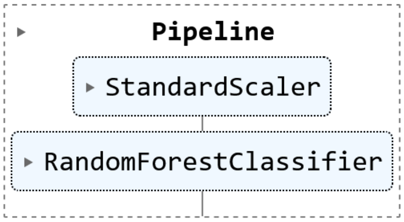
It was also discovered that there were duplicated IDs that leaked into the test set. To better predict the unlabelled data, the duplicated values were deleted from the test set and the new test set only involved 295 unlabelled instances. So, the task involved using the labeled instances to train the models and predict the 295 unlabeled instances.

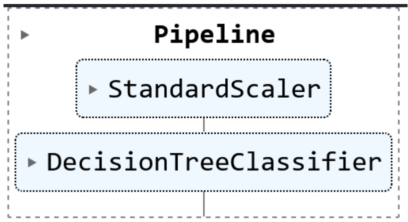
Experiment Design

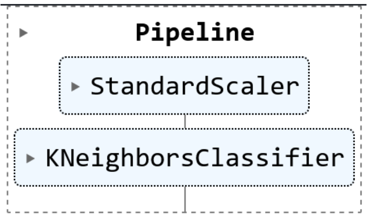
The training set was first split into 80% train and 20% validate sets. Following this, the preliminary assessment of the train set was done using support vector machines with polynomial kernel and RBF kernel. A pipeline was created that included a standard scaler and the SVM classifier The hyperparameters of the polynomial kernel included degree and C where the latter controls the regularization and helps decides the classification boundary. The model was used to fit the X_train and y_train from the train set and trained by trying different hyperparameters. The model was used to predict on X_test and the accuracy score from this prediction was used as the baseline for other models to compare. The evaluation metric for this data set was determined to be the accuracy score on the X_test.



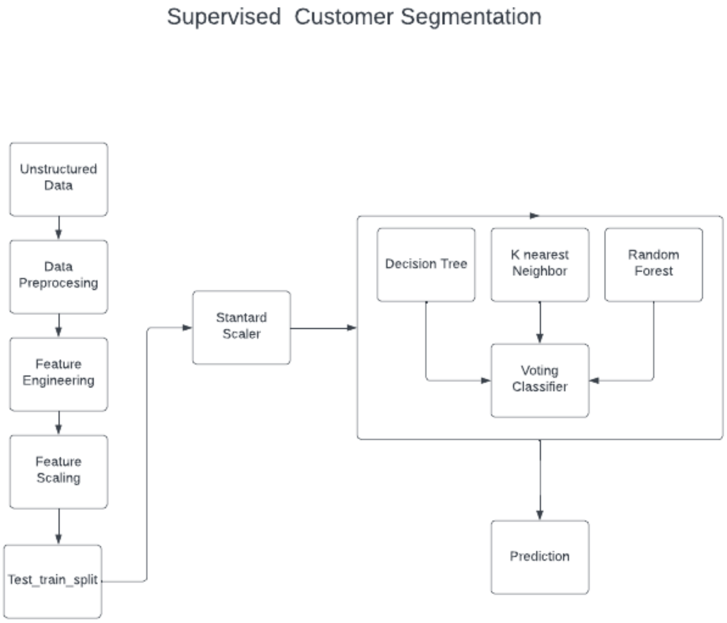
The three algorithms that were chosen after the preliminary assessment was the K-nearest neighbor, Random Forest classifier, and Decision Tree classifier. The same steps that were applied to SVM classifiers were repeated with the three algorithms with different hyperparameters and their accuracy scores were recorded.

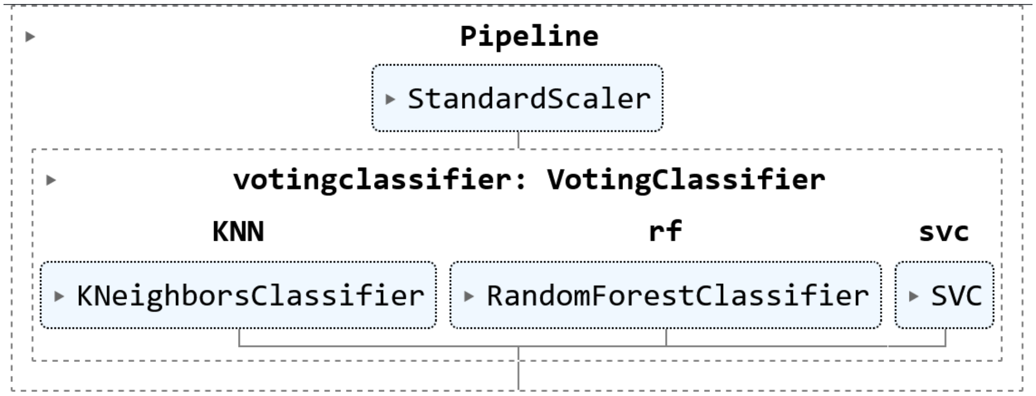






To avoid overfitting, a voting classifier was implemented to vote the predictions on the unlabeled test set. The predictions from each were then voted on and the class that had a majority classification was picked. If all three models returned the same class, the most frequent (mode) was used to make the final prediction.

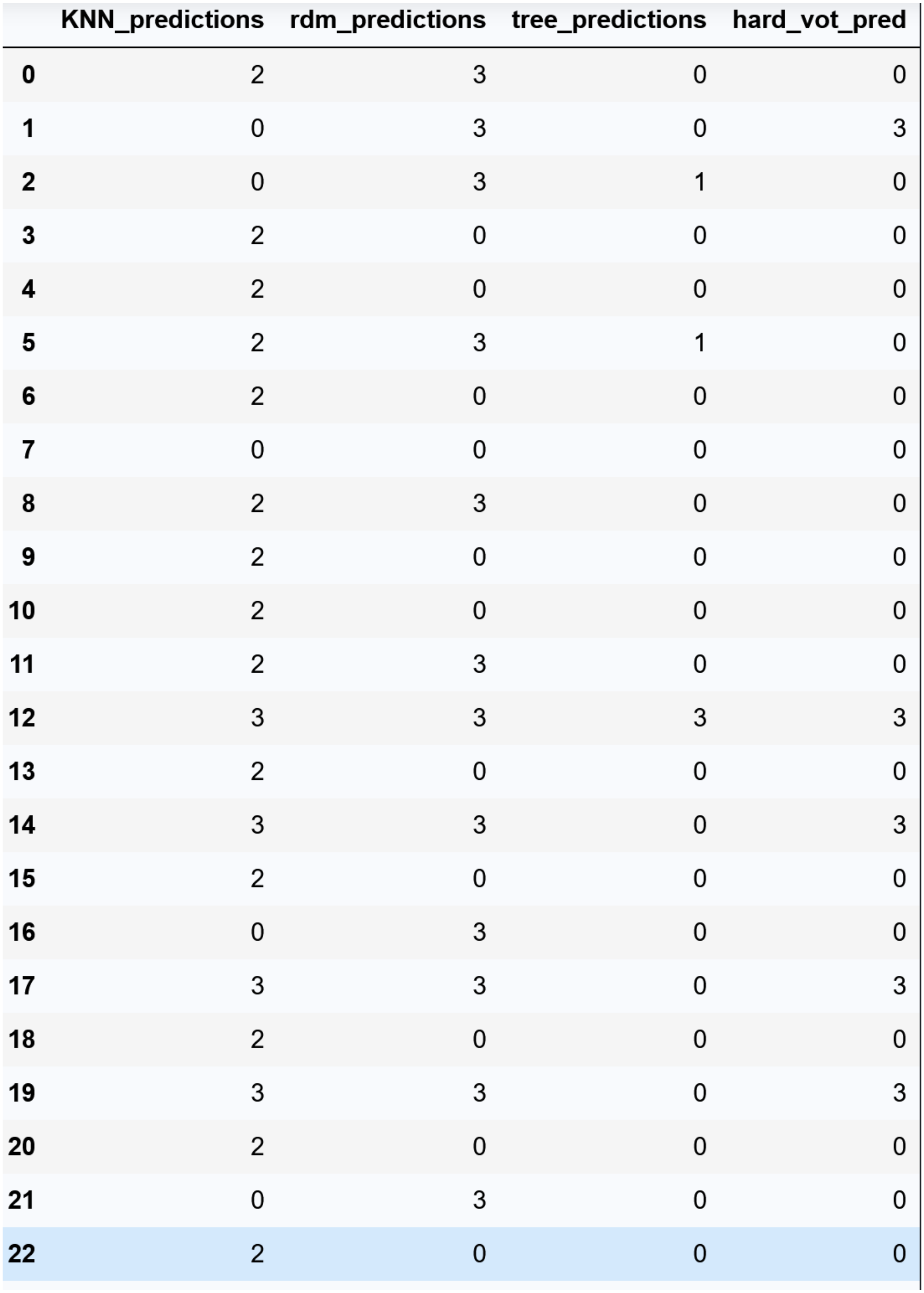




```
▸ Pipeline
▸ StandardScaler
▸ votingclassifier: VotingClassifier
KNN                       rf                          svc
▸ KNeighborsClassifier    ▸ RandomForestClassifier    ▸ SVC
```

Experiment Results

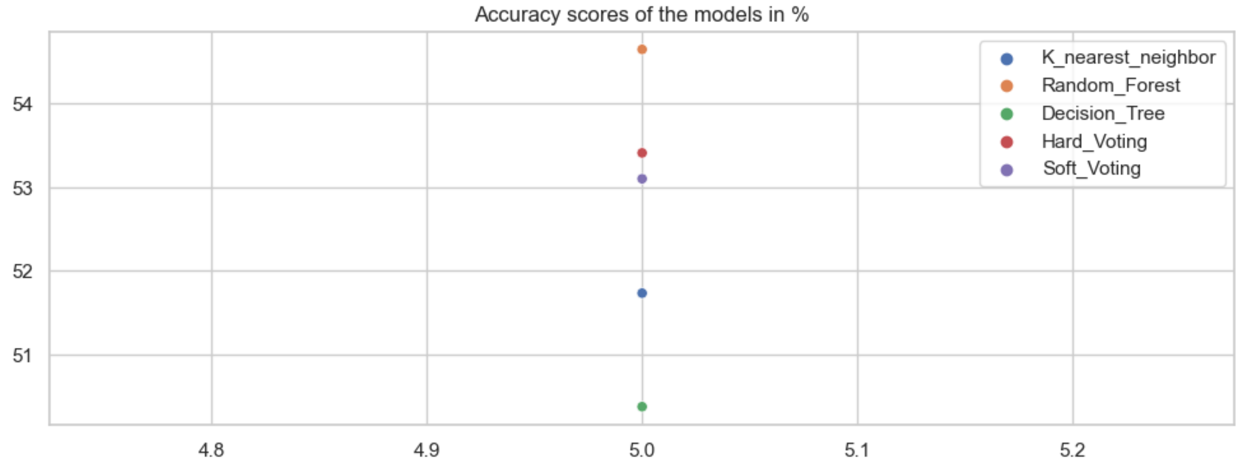
The three classifiers do not differ significantly. They all score between 50% – 55% using the scikit learn's accuracy_score metric. The K neighbors classifier scored 51.73% on the test set. The Random Forest classifier scored 54.65% on the test set while the Decision Tree classifier scored 50.37%. All these models are then fed into a voting classifier which takes the individual predictions from the models and selects the predictions that are made by the majority of the models. A data frame was also created to store the predicted values by all the classifiers and the voting ensemble to compare the results. The voting classifier scored 53.41% using the accuracy_score metric. When all classifiers predict a different value, the voting ensemble uses mode to return a prediction.



	KNN_predictions	rdm_predictions	tree_predictions	hard_vot_pred
0	2	3	0	0
1	0	3	0	3
2	0	3	1	0
3	2	0	0	0
4	2	0	0	0
5	2	3	1	0
6	2	0	0	0
7	0	0	0	0
8	2	3	0	0
9	2	0	0	0
10	2	0	0	0
11	2	3	0	0
12	3	3	3	3
13	2	0	0	0
14	3	3	0	3
15	2	0	0	0
16	0	3	0	0
17	3	3	0	3
18	2	0	0	0
19	3	3	0	3
20	2	0	0	0
21	0	3	0	0
22	2	0	0	0

The hyperparameters can be tuned by using the Gridsearchcv and the Randomizedsearchcv but due to the computational limitation, the best hyperparameters were not produced for optimal performance of the model. A faster GPU would solve this problem. It would also be easier to work with completely unlabelled data because it allows us to use the k means algorithm as discussed in the related works section. However, in this case, the majority of our data is labeled and required us to use the labeled set to train our models and predict unseen unlabeled sets.

The performance of all the models is shown as a scatter plot:



Conclusion

Real-world data is messy and unstructured. It requires extensive preprocessing / cleaning to derive meaningful insights from it. We can use data visualization techniques to understand the distribution of our data and interpret it correctly. While there are many algorithms to choose from, we have to fully understand how the models work to apply them to our given task due to the ethical implications. If we categorize the customers incorrectly, we risk targeting the wrong segment which may be interpreted as unsolicited and unwanted advertising upsetting the mislabelled target group and also exhausting the company resources by using them inefficiently.

References

Kumar, D. (2022, November 14). *Implementing Customer Segmentation Using Machine Learning [Beginners Guide]*. neptune.ai. https://neptune.ai/blog/customer-segmentation-using-machine-learning

Rosas, R. (2021, December 12). *Customer Segmentation using supervised and unsupervised learning*. Medium. https://rrosasl.medium.com/customer-segmentation-using-supervised-and-unsupervised-learning-7522227961ed

Shiral, Y. (2022, June 28). *Customer Segmentation using Machine Learning| Why? | How?* Medium. https://medium.com/@yashashriShiral/customer-segmentation-using-machine-learning-why-how-ffbc3141204f

Real Python. (2022, September 1). *K-Means Clustering in Python: A Practical Guide*. https://realpython.com/k-means-clustering-python/