Grocery Store Optimization

1st Vatsal Jigar Kadakia
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
vkadakia@stevens.edu

2nd Kislay Kislay
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
kkislay@stevens.edu

3rd Mohammed Salman Taqi
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
m4@stevens.edu

***Abstract*—Modern businesses have embraced the use of machine learning algorithms to optimize their operations and improve their overall performance. Machine learning algorithms are used to analyze vast amounts of data to identify patterns, predict outcomes, and make informed decisions. This technology is used in various areas of business, including customer service, marketing, supply chain management, and finance. Our interest specifically focuses on Grocery store management. We chose to implement our learning on how to optimize a grocery store since they have the largest and most frequent number of customers involving major financial aspects as well.**

I. Introduction

We have developed our hypothetical client called “ABC Grocery”. Our client that is the grocery store is in serious need to optimize and improve its performance in several aspects compared to it’s competitors. Different chains of grocery stores are now in fierce competition to become the most famous and profitable brand. For that, they have left behind traditional methods and focuses now more on technical assessments of their company.

Earlier this year, “ABC Grocery” employed a consulting company to provide them “loyalty scores” for their customers. Loyalty score measures the percentage of grocery spend that a customer allocates to “ABC Grocery” vs their competitors. Unfortunately, the consulting company was only able to match around half of the customers to their loyalty database, the other half is missing. Our regression task is to find a way to predict the missing loyalty scores. That would add huge value to their business as it would enable them to target all customers with specific offers and discounts based on their loyalty to ABC Grocery.

Next is the classification challenge. ABC Grocery sent out mailers in a marketing campaign for their “delivery club”. This was a new initiative that cost customers $100 per year for membership, and offered free grocery deliveries rather than the normal cost of $10 per delivery. But sending to everyone proved to be very expensive and next time they would like to save on costs by sending mailers to those customers that are most likely to sign up again. Our classification task is to find out the probability of any customer signing up. Perhaps there are relationships between their shopping behaviour and their characteristics as a customer.

Next task is to implement clustering technique. Senior Management Team of ABC Grocery can’t quite agree on how to think about different types of grocery customers that are shopping with them. Some of the team think that everyone buys from all food departments, just varying amounts depending on how many are in their household. Others in the team think there are specific diets or preferences at play, meaning that some customers shop or don’t shop in certain product areas due to their lifestyles. Our clustering task is that we have to look through the data and try to “segment” up the customers based on any differences observed. If there are different groups within customer base, we have to report back with results. These results will help ABC Grocery to serve their customers in a more targeted way.

The next challenge is Dimensionality Reduction. Principal component analysis(PCA) is used for this task. The grocery store is looking to diversify their business a bit and promote Ed Sheeran’s new album along with their products. They have recently purchased some data around the listening habits of their customers, as well as which customers purchased his last album. Our PCA task is to build a model that would predict the customers who might be interested in his new album. If there is a potential, they could look to purchase the whole data. This step is where we start implementing the unsupervised machine learning algorithms.

Last task is to apply the concept of yet another unsupervised ML algorithm - Association Rule Learning. ABC grocery wants to re-jig the alcohol department within their store. Their customers are often complaining that they can’t find the products they want, and are also wanting recommendations about which other products they should try. The store has provided 3,500 alcohol transactions to deal with this problem.

II. Related Work

”Predicting Product Sales in Retail Stores using Machine Learning” by Yichen Huang, Hui Yang, and Tianqi Wang”. In this paper, the authors use several machine learning algorithms, including support vector regression, decision trees, and random forests, to predict product sales in a Chinese grocery store. ”A Machine Learning Approach to Predict Grocery Sales” by S. S. Mohan and V. S. Raghavan. In this paper, the authors use several machine learning algorithms, including linear regression, decision trees, and neural networks, to predict grocery sales based on historical sales data from a US grocery store. ”Smart Inventory Management in Supermarkets: A Machine Learning Approach” by M. Tharani, S. V. Raja, and M. S. Ramkumar. In this paper, the authors propose a machine learning approach to inventory management in supermarkets, using algorithms such as k-nearest neighbors, decision trees, and neural networks to predict demand for perishable goods in an Indian supermarket.

III. Our Solution

For the first part of the problem statement, we implemented Regression algorithms - Linear, Decision Tree and Random

Forest. Then we built and compared the models using methods and functions of ML libraries(sklearn, matplotlib, numpy, pandas) to predict the missing loyalty scores of customers. Same was implemented for the 2nd part but instead of Regression algorithms, we used Classification techniques namely Decision Tree, Random Forest, Logistic Regression and KNN and compared them to observe which technique was best to find the probability of customers signing up for the delivery club. For the next part, K-means clustering algorithm is used to find the different the types of grocery customers. Clustering helped to find out how many customers belonged to which segment of food or non-food categories. In PCA, we had to predict/find which customers were likely to purchase Ed Sheeran's new album. For that, the predictors were transformed and reduced into components to find the ideal number of components which helped us realize the listening habits of the customers. After an ideal number of components was calculated, the variance that is the variety of listening habits was displayed. For the last task, Association Rule Learning was implemented using Apyori library of python. This library eases the construction of user interests. Identifies the importance of different itemsets. The support function helps to identify different types of importance in itemsets. In our case, its the alcohol department of the grocery store. We had to find what combinations of alcohols were purchased together by the customers so that the grocery stores knows what recommendations of brands to keep together to ease the process of customers which shopping.

A. Description of Dataset

The dataset consists of a grocery_database which consists of 5 individual sheets, various csv files for data extraction and manipulation and pickle files to store the built models and use them later for comparisons. The dataset consists of more than 900 customers of a grocery store including every minute details of transaction dates, sales cost, singup flags, customer ids, loyalty scores etc. Our source of the dataset is Mr. Andrew Jones, who is a well-known data scientist and instructor. He provided real-time datasets from his previous projects. There are 5 csv files along with the main grocery_database. For the regression algorithms, sample_data_regression file was used which has 1 output and 3 input variables. For classification algorithms, sample_data_classification file was used which again has 1 output and 3 input variables. In clustering, sample_data_clustering file was used which has 2 types of variables. For PCA, the csv file consisted of 100 different artists and user_id of all the customers and their listening habit scores. For the last task, the csv file consisted of more than 3,500 transaction id of different alochol brands of more than 49 products.

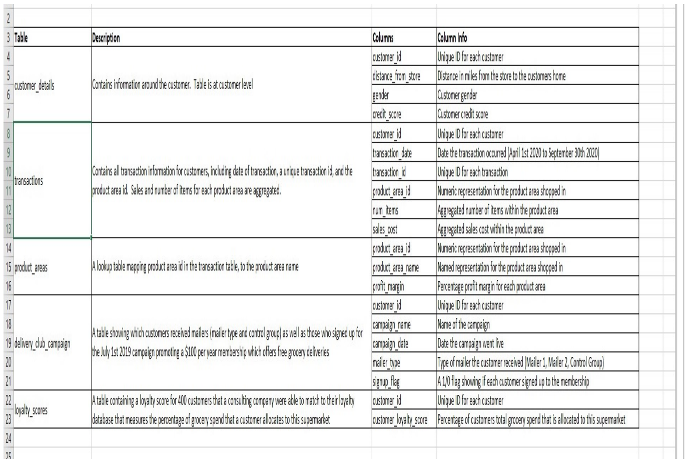
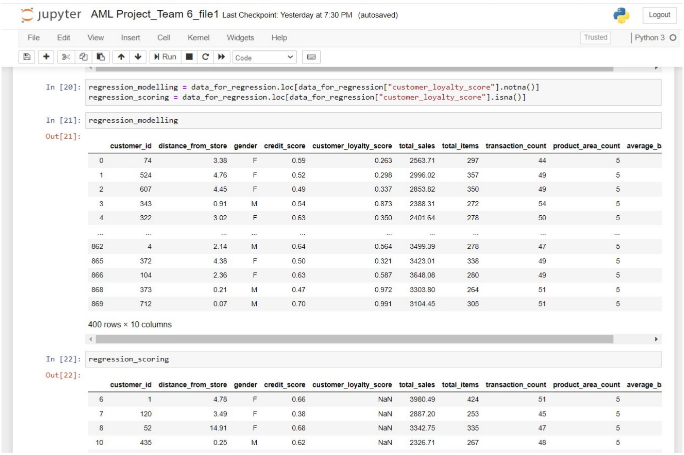


Table	Description	Columns	Column Info
customer_details	Contains information around the customer. Table is at customer level	customer_id	Unique ID for each customer
		distance_from_store	Distance in miles from the store to the customers home
		gender	Customer gender
		credit_score	Customer credit score
transactions	Contains all transaction information for customers, including date of transaction, a unique transaction id, and the product area id. Sales and number of items for each product area are aggregated.	customer_id	Unique ID for each customer
		transaction_date	Date the transaction occurred (April 1st 2020 to September 30th 2020)
		transaction_id	Unique ID for each transaction
		product_area_id	Numeric representation for the product area shopped in
		num_items	Aggregated number of items within the product area
		sales_cost	Aggregated sales cost within the product area
product_areas	A lookup table mapping product area id in the transaction table, to the product area name	product_area_id	Numeric representation for the product area shopped in
		product_area_name	Named representation for the product area shopped in
		profit_margin	Percentage profit margin for each product area
delivery_club_campaign	A table showing which customers received mailers (mailer type and control group) as well as those who signed up for the July 1st 2019 campaign promoting a $100 per year membership which offers free grocery deliveries	customer_id	Unique ID for each customer
		campaign_name	Name of the campaign
		campaign_date	Date the campaign went live
		mailer_type	Type of mailer the customer received (Mailer 1, Mailer 2, Control Group)
		signup_flag	A 1/0 flag showing if each customer signed up to the membership
loyalty_scores	A table containing a loyalty score for 400 customers that a consulting company were able to match to their loyalty database that measures the percentage of grocery spend that a customer allocates to this supermarket	customer_id	Unique ID for each customer
		customer_loyalty_score	Percentage of customers total grocery spend that is allocated to this supermarket

B. Creating the Data

We first read the grocery_database.xlsx and stored the data into variables of loyalty scores, transactions and customer details. Then we created models of data_for_regression, regression_modelling and regression_scoring and stored all dataframes and models into a pickle file using pickle library of python to use the whole models for regressions of different types and classifications. This was the common process for all the algorithms, - Supervised and Unsupervised. Just the main response variables differed with the type of problem statements. In K-means clustering, transactions and product areas dataframes were created instead of loyalty scores to build the clusters. In KNN, it was total sales and total items. In PCA, it was the number of artists and purchased albums.


jupyter
AML Project_Team 6_file1 Last Checkpoint: Yesterday at 7:30 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [20]: regression_modelling = data_for_regression.loc[data_for_regression["customer_loyalty_score"].notna()]
         regression_scoring = data_for_regression.loc[data_for_regression["customer_loyalty_score"].isna()]

In [21]: regression_modelling

Out[21]:
     customer_id  distance_from_store  gender  credit_score  customer_loyalty_score  total_sales  total_items  transaction_count  product_area_count  average_b
  0           74                 3.38       F          0.59                   0.263      2563.71          297                 44                   5
  1          524                 4.76       F          0.52                   0.298      2996.02          357                 49                   5
  2          607                 4.45       F          0.49                   0.337      2853.82          350                 49                   5
  3          343                 0.91       M          0.54                   0.873      2388.31          272                 54                   5
  4          322                 3.02       F          0.63                   0.350      2401.64          278                 50                   5
...          ...                  ...     ...           ...                     ...          ...          ...                ...                 ...
862            4                 2.14       M          0.64                   0.564      3499.39          278                 47                   5
865          372                 4.38       F          0.50                   0.321      3423.01          338                 49                   5
866          104                 2.36       F          0.63                   0.587      3648.08          280                 49                   5
868          373                 0.21       M          0.47                   0.972      3303.80          264                 51                   5
869          712                 0.07       M          0.70                   0.991      3104.45          305                 51                   5

400 rows × 10 columns

In [22]: regression_scoring

Out[22]:
     customer_id  distance_from_store  gender  credit_score  customer_loyalty_score  total_sales  total_items  transaction_count  product_area_count  average_b
  6            1                 4.78       F          0.66                     NaN      3980.49          424                 51                   5
  7          120                 3.49       F          0.38                     NaN      2887.20          253                 45                   5
  8           52                14.91       F          0.68                     NaN      3342.75          335                 47                   5
 10          435                 0.25       M          0.62                     NaN      2326.71          267                 48                   5
```

C. Machine Learning Algorithms

To determine factors behind the missing loyalty scores of customers, we used Linear Regression as the first case. In Linear regression, it is very important to deal with the outliers, then split input and output variables. Another important factor is to drop the categorical variables and assign them binary values. Next we selected the feature. We used Recursive Feature Elimination with Cross-validation also known as RFECV from scikit library. Its parameters and attributes are very useful for estimating and calculating results. Finally we created the data model, dropped the missing values and trained, tested, split the data to calculate r-squared score, cv(cross validation) score and adjusted the r squared score.

Next, we carried out the same process for Decision Tree Regression but with additional features.A Decision Tree is a model that splits the data into distinct buckets using input variables, with split decisions being based on how well each potential split explains differences in the output variable. Over fitting is the most common error cause in Decision Trees. We demonstrated and took care of over fitting by calculating the maximum depth of the list of features to get an approximation of the accuracy of our model and compare it with model of Linear Regression. Lastly for the final comparison in regression, we implemented the same model using Random Forest algorithm. A Random Forest is an ensemble model consisting of many Decision Trees working together across different randomly selected subsets of the data, facilitating improved accuracy and stability. In RF, feature_importance and permutation_importance are displayed and calculated using the trained and tested data.

For the Classification task, to find the probability of customers signing up for the delivery club we used all the classification algorithms(Decision Tree, RF, Logistic, KNN) to determine the difference in the confusion matrices, f1 score, precision and recall scores. These all parameters decide why and what type of customers decide whether to signup for delivery club or not. In Logistic Regression classifier, what stands out is finding the optimal threshold of classification model.

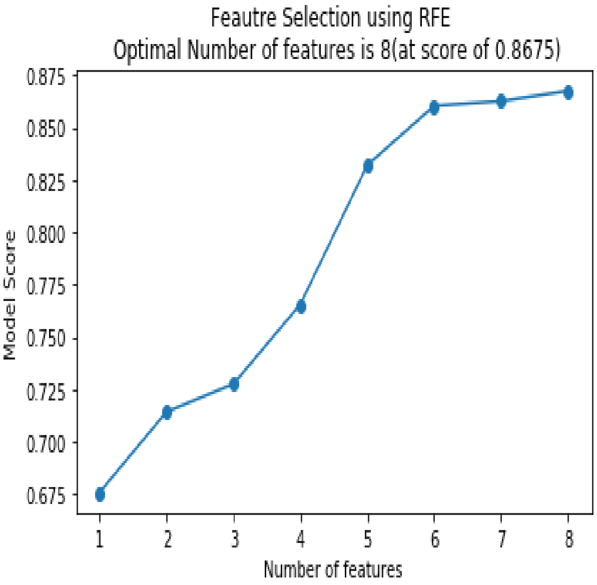
For the next task, K-means Clustering is used as we can clearly observe the cluster of customers by aggregating the data and then dropping the non-food category. The algorithm created four clusters - Dairy, Fruit, Meat and Vegetables and distributed the customers in respective clusters based on their customer_id, transaction summaries, sales of each category and product areas.

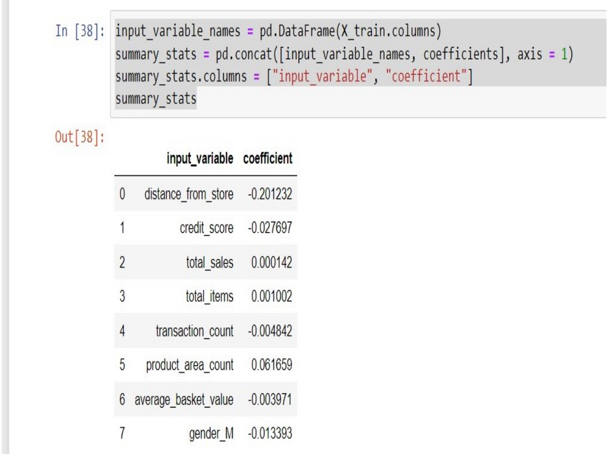
For the next two tasks, our focus shifts on analysing data using unsupervised ML algorithms - Principal Component Analysis and Association Rule Learning. The goal is to find structures and patterns from the provided data. PCA is often used as a Dimensionality Reduction technique that can reduce a large set of variables down to a smaller set that still contains most of the original information. We blend the original features/ variables and create components which has the key information of the dataset. For this, just like k-means we have to manually input how many number of principal components we want. Before applying PCA, it is important to to always apply scaling to your variables. We used standardisation instead of normalisation. It is bound to lose some of the information/variance contained in original data. Lastly, it will be much harder to interpret the outputs based on component values versus the original variables. So in our case there are 100 artists. The listening habits of customers were boiled down to 20-25 components including all the artists and determined if the customers will purchase the new album or not.

Association Rule Learning is an approach that discovers the strength of relationships between different data points. It is commonly utilised to understand which products are frequently purchased together. In this task, we had to find what brands of alcohols were purchased together. For example, French wines - red and white were frequently purchased together. From this we can hypothesize the purchasing habits from the transactions. Apriori algorithm is used for this which is the most part of this task. It is divided into 4 key areas - a) Support - percentage of all transactions that contain both item A and item B. b) Confidence - of all transactions that included item A, what proportion also included item B. c) Expected Confidence - percentage of all transactions that contain item B. d) Lift - The factor by which the Confidence exceeds the Expected Confidence. This all factors help the grocery store how to present or arrange alcohol brands and types together. Understand if discounts should be offered on one of the items. Understand if only one of the alcohol items should be advertised at a time,

D. Implementation Details

1) Linear regression graph below indicated the number of optimal features is 8 at the maximum score of fit.grade_scores 0.8675. Model training is done on train set and model assessment is done on test set. Test size is 20%. After calculating r-squared score(78 %) and cv score(85%) we extract the model coefficients and get to know that the distance from the grocery store to customer's house was the most affected factor in determining the loyalty score. As we can observe from the table of coefficients, distance from the store has the most negative coefficient meaning the customers are less likely to come to ABC grocery if the distance is too far away from their houses.



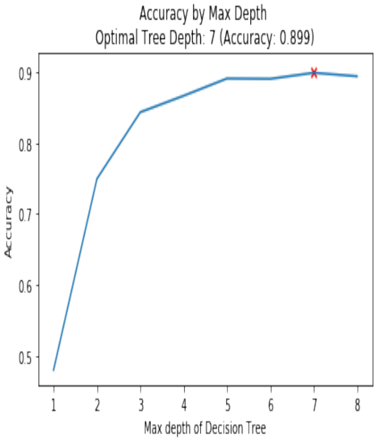


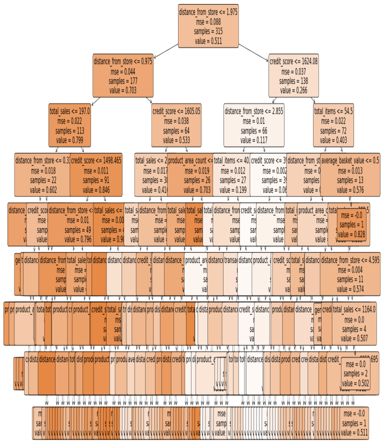
```
In [38]: input_variable_names = pd.DataFrame(X_train.columns)
         summary_stats = pd.concat([input_variable_names, coefficients], axis = 1)
         summary_stats.columns = ["input_variable", "coefficient"]
         summary_stats
```

Out[38]:

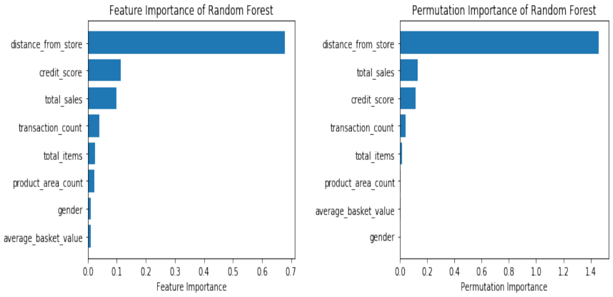
	input_variable	coefficient
0	distance_from_store	-0.201232
1	credit_score	-0.027697
2	total_sales	0.000142
3	total_items	0.001002
4	transaction_count	-0.004842
5	product_area_count	0.061659
6	average_basket_value	-0.003971
7	gender_M	-0.013393

2)The implementation graphs for decision trees are displayed below. They also indicate the same that distance_from_store is the most affecting factor. Demonstration of over fitting is done using regressor.predict(X_train). Then the best maximum depth for the optimal tree is calculated and decision tree is plotted. The maximum depth of the tree(also the optimal depth) turned out to be 7 with root node as distance from the store. Before dealing with over fitting, r-squared score was approximately 86% and after adjusting the over fitting it improved to around 90%. Overall, the accuracy of the model was almost 90%. Cross validation score was 80% with k=4 splits.

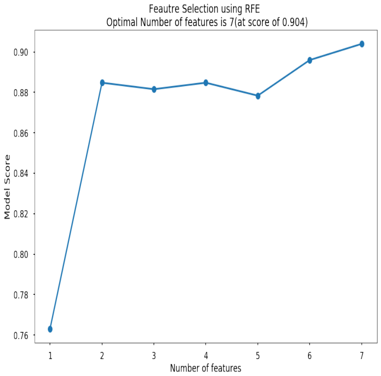


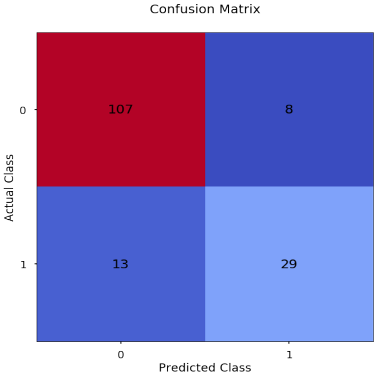


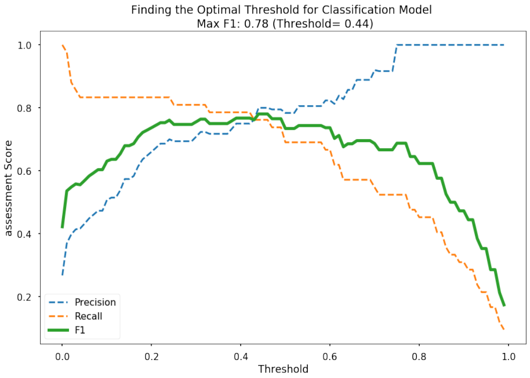
3)In Random Forest Regression, the bar graphs of selected features and permutation confirm our result that distance from the grocery store matters a lot to maintain loyalty of customers. Distance from the store has the highest feature importance of 0.7 and permutation importance of 1.4. Other variables have significantly low importance in the decision trees. In regression, this was by far the best performing algorithm compared to linear and decision tree with a r-squared score of 96%, cross validation score of 92%, adjusted r-squared score of approximately 95%. An array of predictions was displayed in the code for all the decision trees.



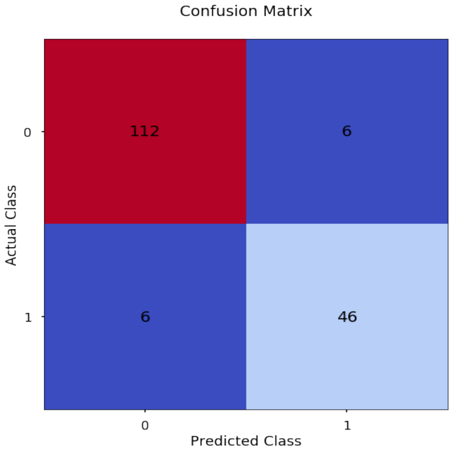
4) Logistic Regression is a model used to predict the probability of a certain event or class based one or more input variables. It transforms linear relationships to a probabilistic output through the Logistic Function. Although it is a regression model, it is widely used for classification problem. The probability for the customer signup rate is calculated for the delivery club campaign. From the confusion matrix, we can observe that 107 are True Positive values meaning most of our prediction was correct compared to the True Negative and False Positive values of 29 and 13 respectively. Accuracy score(number of correct classification out of all attempted classifications) is 86%. Precision score(of all observations that were predicted as positive, how many are actually positive) of 78%. Recall score(of all positive observations how many did we predict as positive) of 69% and f1 score (the harmonic mean of precision and recall) of 73%. Optimal number of features is 7 with an optimal threshold of 0.44.

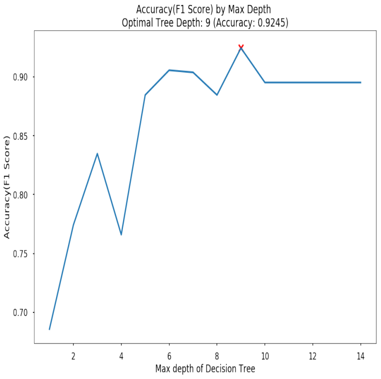


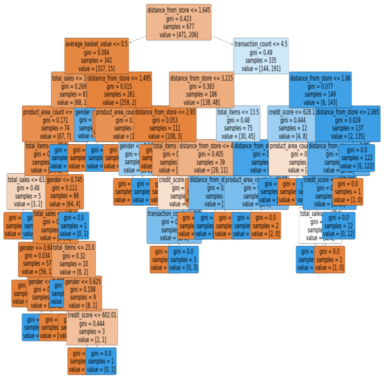




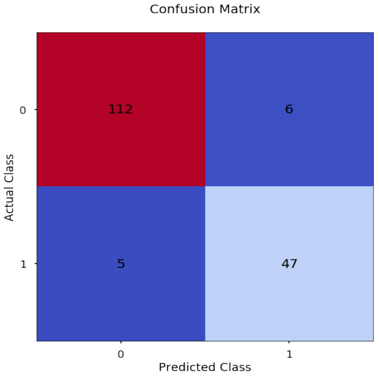
5) In Decision Tree Classification, optimal depth of the tree is 9. In regression it was 7. Hence, this tree classifies more leaf nodes. Maximum model accuracy is 92%. Precision score of 88%, Recall score of 88% as well and f1 score of 88% as well. There are 112 True Positive values and 46 True negative values indicating that the model can be much better. But its a little better than Logistic regression classification.

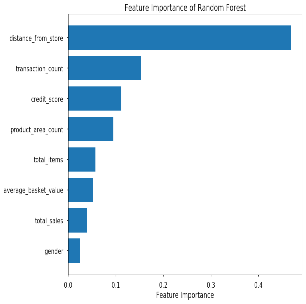


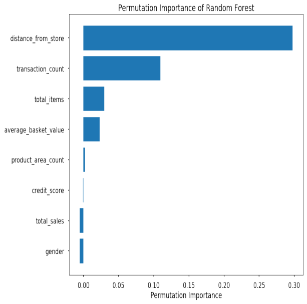




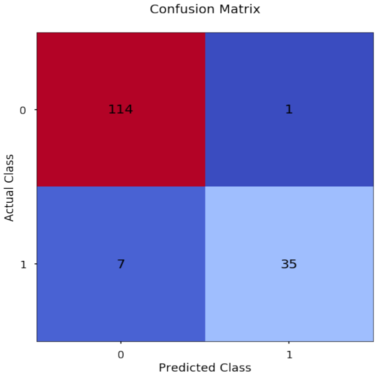
6)In Random Forest classification, model accuracy is 93%. Precision score is 88%. Recall score is 90% and f1 score is 89%. The confusion matrix is almost similar to Decision tree classifier, just one less False Positive value which lead to a better model than decision tree. So in both cases, Random Forest turned out to be the best algorithm for this particular grocery dataset.

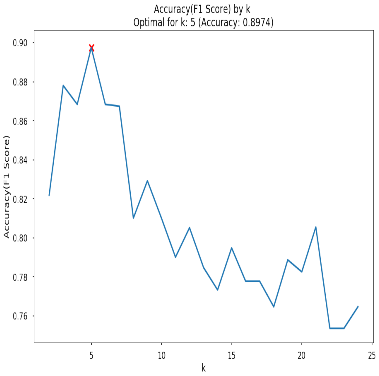




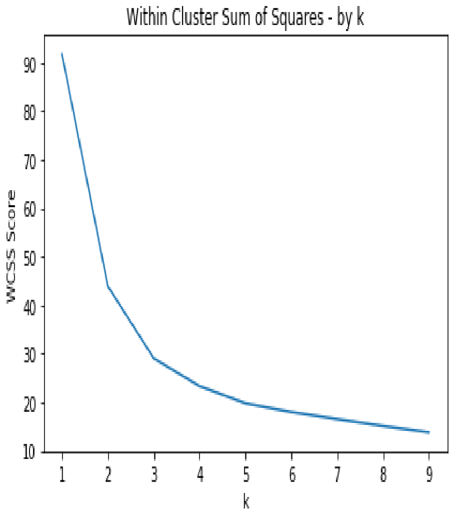


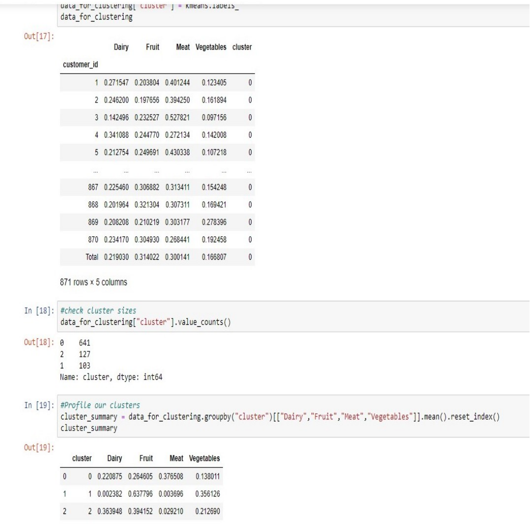
7)The KNN algorithm predicts a class for an unknown data point using the most popular class of a number of nearby known data points. The number of nearby data points used to form a prediction is denoted by k. The feature selection in this case is Random Forest classifier. Model accuracy is 95%. Precision score of 97%. Recall score of 83% and f1 score of 89%.. Optimal value for k obtained was 5. The confusion matrix is comparatively much better with 114 True Positive values.





8) K means Clustering partitions data points into distinct groups based on similarity with each other. the number of distinct groups is determined by k. Here sales are aggregated at customer level by product area to accurately predict which customer falls into which category. Since Food category is highly relevant than non-food category in this dataset to determine the sales and purchasing habits of customers, non food category is dropped. 3 distinct clusters are created in a dataframe as displayed below in the figure. Data is normalized using MinMaxScaler method.





```
data_for_clustering["cluster"] = kmeans.labels_
data_for_clustering
```

Out[17]:

customer_id	Dairy	Fruit	Meat	Vegetables	cluster
1	0.271547	0.203804	0.401244	0.123405	0
2	0.246200	0.197656	0.394250	0.161894	0
3	0.142496	0.232527	0.527821	0.097156	0
4	0.341088	0.244770	0.272134	0.142008	0
5	0.212754	0.249691	0.430338	0.107218	0
...	...	...	...	...	...
867	0.225460	0.306882	0.313411	0.154248	0
868	0.201964	0.321304	0.307311	0.169421	0
869	0.208208	0.210219	0.303177	0.278396	0
870	0.234170	0.304930	0.268441	0.192458	0
Total	0.219030	0.314022	0.300141	0.166807	0

871 rows × 5 columns

In [18]:

```
#check cluster sizes
data_for_clustering["cluster"].value_counts()
```

Out[18]:

```
0    641
2    127
1    103
Name: cluster, dtype: int64
```

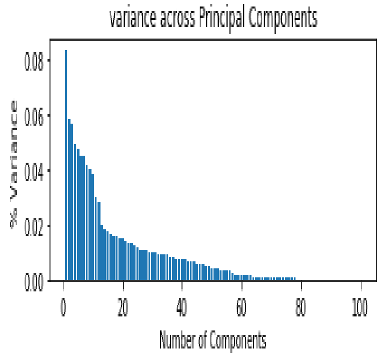
In [19]:

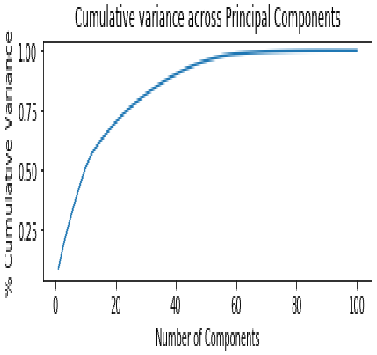
```
#Profile our clusters
cluster_summary = data_for_clustering.groupby("cluster")[["Dairy","Fruit","Meat","Vegetables"]].mean().reset_index()
cluster_summary
```

Out[19]:

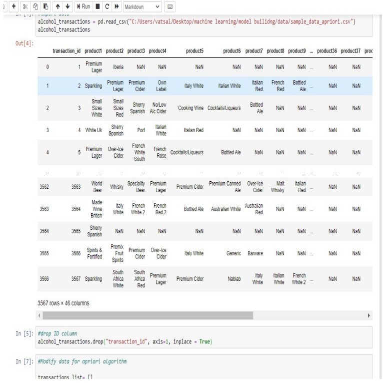
	cluster	Dairy	Fruit	Meat	Vegetables
0	0	0.220875	0.264605	0.376508	0.138011
1	1	0.002382	0.637796	0.003696	0.356126
2	2	0.363948	0.394152	0.029210	0.212690

9)In Principal Component Analysis, StandardScaler method is used to scale the data into components. First the variance of each component is displayed and then the cumulative variance graph is plotted. 0.75% cumulative variance is obtained and with the help of that we can observe that 100 different artists are divided into 20-25 components. Hence the data is scaled down by including all the artists to determine the listening habits of customers and decide whether they will purchase the new album of Ed Sheeran or not. Model accuracy is calculated taking Random Forest classifier since it was the best performing algorithm in our case. 92% accuracy was achieved which is very good.





10)In Association Rule Learning, missing values of all the 3,500 alcohol transactions are first dropped. Then using apriori algorithm, support, confidence and lift values are calculated and displayed. The combinations of various alcohol brands are displayed. For example, we can observe that Wine Gifts and Beer gifts are most frequently purchased together as a combination. So the customer either buys wine gifts or beer gifts and they should be arranged together for the better comfort of customers.



```
alcohol_transactions = pd.read_csv("C:/Users/vatsal/Desktop/machine learning/model builidng/data/sample_data_apriori.csv")
alcohol_transactions
```

Out[4]:

	transaction_id	product1	product2	product3	product4	product5	product6	product7	product8	product9	...	product36	product37	proc
0	1	Premium Lager	Iberia	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
1	2	Sparkling	Premium Lager	Premium Cider	Own Label	Italy White	Italian White	Italian Red	French Red	Bottled Ale	...	NaN	NaN	
2	3	Small Sizes White	Small Sizes Red	Sherry Spanish	No/Low Alc Cider	Cooking Wine	Cocktails/Liqueurs	Bottled Ale	NaN	NaN	...	NaN	NaN	
3	4	White Uk	Sherry Spanish	Port	Italian White	Italian Red	NaN	NaN	NaN	NaN	...	NaN	NaN	
4	5	Premium Lager	Over-Ice Cider	French White South	French Rose	Cocktails/Liqueurs	Bottled Ale	NaN	NaN	NaN	...	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3562	3563	World Beer	Whisky	Speciality Beer	Premium Lager	Premium Cider	Premium Canned Ale	Over-Ice Cider	Malt Whisky	Italian Red	...	NaN	NaN	
3563	3564	Made Wine British	Italy White	French White 2	French Red 2	Bottled Ale	Australian White	Australian Red	NaN	NaN	...	NaN	NaN	
3564	3565	Sherry Spanish	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
3565	3566	Spirits & Fortified	Premix Fruit Spirits	Premium Cider	Over-Ice Cider	Italy White	Generic	Barware	NaN	NaN	...	NaN	NaN	
3566	3567	Sparkling	South Africa White	South Africa Red	Premium Lager	Premium Cider	Nablab	Italy White	Italian White	French White 2	...	NaN	NaN	

3567 rows × 46 columns

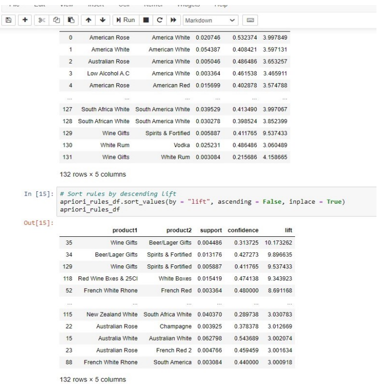
In [5]:

```
#drop ID column
alcohol_transactions.drop("transaction_id", axis=1, inplace = True)
```

In [7]:

```
#Modify data for apriori algorithm

transactions_list= []
```



0	American Rose	America White	0.020746	0.532374	3.997849
1	America White	American White	0.054387	0.408421	3.597131
2	Australian Rose	America White	0.005046	0.486486	3.653257
3	Low Alcohol A.C	America White	0.003364	0.461538	3.465911
4	American Rose	American Red	0.015699	0.402878	3.574788
...	...	...	...	...	...
127	South Africa White	South America White	0.039529	0.413490	3.997067
128	South African White	South America White	0.030278	0.398524	3.852399
129	Wine Gifts	Spirits & Fortified	0.005887	0.411765	9.537433
130	White Rum	Vodka	0.025231	0.486486	3.060489
131	Wine Gifts	White Rum	0.003084	0.215686	4.158665

132 rows × 5 columns

```
In [15]: # Sort rules by descending lift
         apriori_rules_df.sort_values(by = "lift", ascending = False, inplace = True)
         apriori_rules_df
```

Out[15]:

	product1	product2	support	confidence	lift
35	Wine Gifts	Beer/Lager Gifts	0.004486	0.313725	10.173262
34	Beer/Lager Gifts	Spirits & Fortified	0.013176	0.427273	9.896635
129	Wine Gifts	Spirits & Fortified	0.005887	0.411765	9.537433
118	Red Wine Bxes & 25Cl	White Boxes	0.015419	0.474138	9.343923
52	French White Rhone	French Red	0.003364	0.480000	8.691168
...	...	...	...	...	...
115	New Zealand White	South Africa White	0.040370	0.289738	3.030783
22	Australian Rose	Champagne	0.003925	0.378378	3.012669
15	Australia White	Australian White	0.062798	0.543689	3.002074
23	Australian Rose	French Red 2	0.004766	0.459459	3.001634
88	French White Rhone	South America	0.003084	0.440000	3.000918

132 rows × 5 columns

IV. Comparison

For 1st task in predicting the loyalty score and determining it's factors, Random Forest Regression was the best algorithm as it predicted the best r2 score, cv score and adjusted r2 score. Its bar graphs also demonstrated very clearly and precise that distance from the store was the main factor. All of the regression and classification algorithms determined distance from store was the most affecting factor in customer loyalty scores, but overall the best models were of Random Forest regression and Random Forest Classifier. The initial process of dropping missing values, joining the tables and aggregating and then dropping the categorical variable using OneHotEncoder which is 'gender' in this dataset, split the data into training and testing is the same for all algorithms. Then the later feature selection differs. Data was normalized or scaled using various methods like MinMaxScaler, StandardScaler etc in different algorithms. For 2nd task, KNN classifier stands out of the 4 algorithms because it had better accuracy score, precision, recall and f1 score all of them near or above 90%. Association rule learning and PCA were distinct algorithms used for different problem statements and there is no real comparison in that. But both used Random forest classification since it was the best one in our cases.

V. Future Directions

We missed out on implementing Causal Impact Analysis. Since it is a time series analysis concept, at this moment we weren't equipped enough to deal with it. That was the only algorithm we missed out on and would like to work on it in future. Since we offered real time ML solutions, we would like to convert our project in form of a product and publish this so that it gets noticed and we as a group can offer data analysis services to commercial businesses.

VI. Conclusion

One significant way machine learning algorithms have transformed businesses is by providing more personalized and efficient customer service. By analyzing customer data, machine learning algorithms can identify patterns in customer behavior and preferences, enabling businesses to tailor their products and services to individual customers. This not only enhances the customer experience but also increases customer retention and loyalty. Another area where machine learning algorithms have proven useful is in marketing. By analyzing customer data, businesses can identify potential customers, target them with personalized advertising, and predict their behavior, allowing them to make informed decisions about their marketing strategy. Overall, machine learning algorithms have become an essential tool for modern businesses to improve their performance, reduce costs, and gain a competitive advantage in the market.

References

1) ”Predicting Product Sales in Retail Stores using Machine Learning” by Yichen Huang, Hui Yang, and Tianqi Wang. https://ieeexplore.ieee.org/document/8745689
2) ”A Machine Learning Approach to Predict Grocery Sales” by S. S. Mohan and V. S. Raghavan. https://ieeexplore.ieee.org/document/8373837
3) ”Smart Inventory Management in Supermarkets: A Machine Learning Approach” by M. Tharani, S. V. Raja, and M. S. Ramkumar. https://ieeexplore.ieee.org/document/8253466