**CUSTOMER ACQUISTION COST PREDICTION**

Contents

[PROJECT DESCRIPTION 2](#_Toc116235595)

[DATA EXPLORATION AND PREPROCESSING: 2](#_Toc116235596)

[ONE – HOT ENCODING 4](#_Toc116235597)

[Linear Regression: 4](#_Toc116235598)

[Random Forest: 4](#_Toc116235599)

[Decision Tree Regressor: 4](#_Toc116235600)

[kNN 5](#_Toc116235601)

[ORDINAL ENCODING 6](#_Toc116235602)

[Linear Regression: 6](#_Toc116235603)

[Ridge Regression: 6](#_Toc116235604)

[Lasso Regression: 7](#_Toc116235605)

[Random Forest: 7](#_Toc116235606)

[Decision Tree Regressor: 8](#_Toc116235607)

[kNN: 8](#_Toc116235608)

[Neural Network: 9](#_Toc116235609)

[ENSEMBLE TECHNIQUES 9](#_Toc116235610)

[Random Forest Bagging: 9](#_Toc116235611)

[Random Forest Boosting & XG Boosting: 9](#_Toc116235612)

[Ensemble using Averaging: 10](#_Toc116235613)

[RESULTS AND SUMMARY 11](#_Toc116235614)

[REFERENCES 12](#_Toc116235615)

[GROUP MEMBERS 12](#_Toc116235616)

# PROJECT DESCRIPTION

* We selected a dataset which has sales data for different stores across various states around USA and countries like Mexico and Canada too, the dataset contains various food items and food department in a supermarket.
* We are predicting Cost of customer acquisition from the dataset, through the process of data cleaning and modelling of various types.
* We created a few models in the project report below that contains KNN, Linear Regression, Random Forest, Decision tree regressor.

# DATA EXPLORATION AND PREPROCESSING:

* The output variable for this data analysis is Cost of customer acquisition and the input variables are columns with correlation between -0.8 to 0.8.
* We plotted the mean of the cost of acquisition against various data points like recyclable package, low\_fat, video store, coffee bar, and prepared food. To analyze how much of a difference these data points make in the model's predictability.

Shape, square

Description automatically generatedShape

Description automatically generated

**A picture containing square

Description automatically generatedChart, bar chart

Description automatically generatedPlots for country sales and sales between various states VS the cost of acquisition:**

**Chart, bar chart

Description automatically generated**

**Sales across various USA states vs Cost**

* We could see across the plots that Cost of acquisition when plotted across the various categories has little or almost negligible difference thus, we choose to drop these columns going forward for our further analysis.

**Correlation Matrix for various categories:**

* We have selected the columns that are between the correlation of -0.8 to 0.8 against the Cost of acquisition.

**A computer screen capture

Description automatically generated with medium confidence**

## ONE – HOT ENCODING

### Linear Regression:

Chart, scatter chart

Description automatically generatedText, table

Description automatically generatedTo begin with, the linear regression model is run to understand the significant predictors in the dataset. This is done by observing the p-value of the estimators. With this model, we tried to understand which predictors contributed more towards the performance.

### Random Forest:

Text

Description automatically generated with low confidenceSince RF modeling replaces missing values with majority values, we used the Random Forest model to achieve a good accuracy score.

A picture containing chart

Description automatically generated

### Decision Tree Regressor:

Decision trees can handle continuous and categorical variables and decide which predictors are more critical than others. Also, a decision tree can handle missing values. Hence, we used a decision tree to train and generate our model.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

### kNN

**Graphical user interface, text, application

Description automatically generated**Chart, scatter chart

Description automatically generatedWe tried to use kNN modeling to compete with the most accurate models because it makes precise highly predictions. Our dataset was comparatively huge, which helped us reduce the model’s training period.

**Text

Description automatically generated**

## ORDINAL ENCODING

### Linear Regression:

To begin with, the linear regression model is ran by us to understand the significant predictors in the dataset. This is done by observing the p-value of the estimators. With this model, we tried to understand which predictors contributed more towards the performance

A screenshot of a computer

Description automatically generated with low confidenceText

Description automatically generatedText

Description automatically generated

### Ridge Regression:

Text

Description automatically generatedChart, scatter chart

Description automatically generatedWe have performed Ridge regression to evenly distribute the weights to reduce standard error.

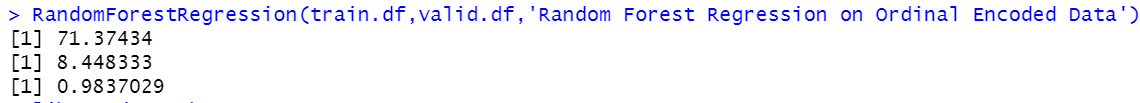
### Lasso Regression:

**Text

Description automatically generated**Chart, scatter chart

Description automatically generatedWe have implemented Lasso regression to eliminate those predictors who do not contribute significantly to the outcome. In Lasso regression, the coefficients of less significant predictors will be set to 0 hence removing them.

### Random Forest:

Chart, scatter chart

Description automatically generatedSince RF modeling replaces missing values with majority values, we used the Random Forest model to achieve a good accuracy score. 

### Decision Tree Regressor:

**Text, letter

Description automatically generated**Chart, scatter chart

Description automatically generatedDecision trees can handle continuous and categorical variables and decide which predictors are more critical than others. Also, a decision tree can handle missing values. Hence, we used a decision tree to train and generate our model.

### kNN:

We tried to use kNN modeling to compete with the most accurate models because it makes precise highly predictions. Our dataset was comparatively huge, which helped us reduce the model’s training period.

Chart, scatter chart

Description automatically generatedText

Description automatically generated

### Neural Network:

Diagram

Description automatically generatedWe have used neural network with 1 hidden layer and also tried increasing the hidden layers to 2. Due to the big size of the data, neural network took a lot of time to train.

We are working towards the rescaling and getting the results from the neural network, this will be one of the parts of our future works.

# ENSEMBLE TECHNIQUES

## Random Forest Bagging:

We used bagging to raise the stability of models in improving accuracy and reducing variance, which eliminates the challenge of overfitting.

**Graphical user interface, text, application

Description automatically generated**Chart, scatter chart

Description automatically generated

## Random Forest Boosting & XG Boosting:

Boosted Regression Trees are a robust algorithm that works very well with large datasets or when one has many environmental variables compared to the number of observations. They are very robust to missing values and outliers.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedXGBoost is a tree-based ensemble machine learning algorithm with higher predicting power, and we used it to improvise our performance. You can see the results above.

**Graphical user interface, text

Description automatically generated with medium confidence**Graphical user interface, text

Description automatically generated

## Ensemble using Averaging:

Graphical user interface, text, application

Description automatically generatedAs averaging ensemble suggests, we used this technique to make better predictions and achieve better performance by taking the average of all the models altogether than any single contributing model. And we can see that it gave a comparatively better R2 score than any other single models above.

Chart, scatter chart

Description automatically generated

# RESULTS AND SUMMARY

In the table below we can find the performance of various models trained on One Hot encoded data which has 110 columns. We can see that best R2 score is achieved by Linear Regression Model. To deal with high dimensionality we also tried PCA but there was not much improvement in results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R2** |
| Linear Regression | 19.8 | 24.2 | 0.351 |
| Random Forest Regressor | 79.5 | 101 | 0.131 |
| DecisionTree Regressor | 20.8 | 25.5 | 0.28 |
| kNN | 22.55 | 28.1 | 0.154 |

In the table below we can find the performance of various models trained on Ordinal encoded data which has 19 columns. We can see that good R2 score is achieved by below models.

* Random Forest Regressor
* XGboosting on Random Forest,
* Baggin on Random Forest
* Boosting on Random Forest
* KNN
* Ensemble with Averaging

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R2** |
| Linear Regression | 25.5 | 5.05 | 0.0356 |
| Random Forest Regressor | 71.4 | 8.44 | 0.985 |
| DecisionTree Regressor | 21.1 | 4.58 | 0.264 |
| Ridge Regression | 25.5 | 5.05 | 0.0355 |
| Lasso Regression | 25.5 | 5.05 | 0.0356 |
| KNN | 5.76 | 13.11 | 0.806 |
| Bagging on RF | 0.193 | 8.19 | 0.998 |
| Boosting on RF | 8.512 | 8.19 | 0.903 |
| XGBoostng on RF | 1.15 | 8.19 | 0.99 |
| Ensemble with Averaging | 14.07 | 16.73 | 0.858 |

# REFERENCES

* <https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart>
* <https://www.geeksforgeeks.org/how-neural-networks-are-used-for-regression-in-r-programming/>

# GROUP MEMBERS

1. Anurag
2. Arunita Sarkar
3. Ashutosh Mishra
4. Nikhil Shenoy
5. Soumyata Jena