VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Smart Internz

APPLIED DATA SCIENCE PROJECT.

Prediction Of Customer Acquisition Cost using Machine Learning.

Anusha Garg (20BIT0255)

Bhavya Nagpal (20BIT0215)

Saatvik Gupta (20BBS0065)

Gursehaj Singh(20BB0093)

INTRODUCTION

1.1 Overview

Customer acquisition can be quite a challenging task in today's world, mostly due to the fact that businesses face intense competitions from rival companies for the same target audience. To fight the competition, they have to formulate marketing strategies and tactics in order to stand out and attract customers. Besides, it can be quite challenging to entice a potential customer to buy your product from a wide range of choice offered to him / her.

This project aims to use Machine Learning algorithms to predict the cost of acquiring a new customer based on the data of about 60,000 existing customers based on multiple factors such as the products sold, store features, income of the customer, etc.

1.2 Purpose

The purpose of this project is to analyse existing consumer behaviour and understand what appeals to them and on the basis of this information gathered, predict the cost of acquiring new customers.

LITERATURE SURVEY

2.1 Existing Problem

Machine learning can be used to predict customer acquisition by analysing historical data and identifying patterns and trends that are indicative of potential new customers. By training machine learning models on past customer acquisition data, these models can learn to recognize the characteristics and behaviours of customers who are likely to be acquired in the future.

[1] Schröder (2023) emphasizes the importance of holistic branch design in acquiring new customers and fostering lasting loyalty. [2] Camacho-Vallejo et al. (2023) proposes a hierarchized green supply chain framework integrating sustainability considerations with customer selection, routing, and nearshoring. [3] Calder et al. (2023) investigate the relationship between customer equity and brand equity, providing insights into the financial value of marketing. [4] Cherchye et al. (2023) present a DEA-based approach for customer value analysis, enabling comparisons across entities. [5] Harsh et al. (2023) introduce an intelligent email categorization system using SVM to enhance customer support efficiency in the pharmaceutical domain.

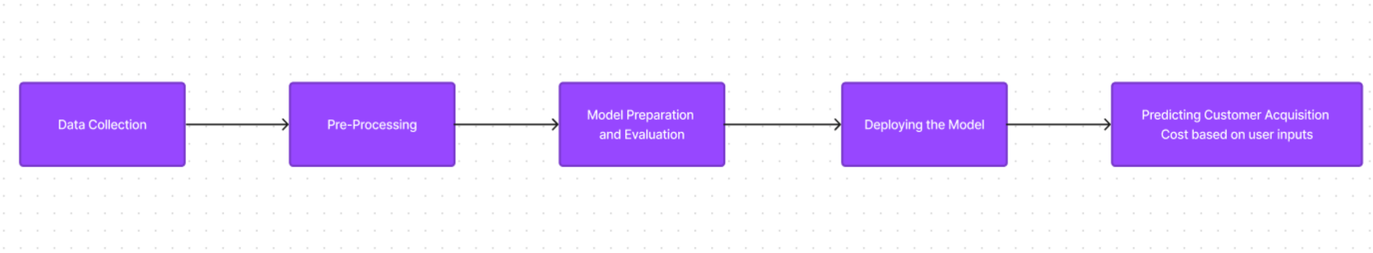
After reviewing multiple papers, we can conclude that there are several challenges when it comes to predicting customer acquisition costs. For example, such predictions hugely rely on the quality of data used for making the predictions and acquiring such data can be quite a task. Secondly, different organisations may use different parameters to estimate their customer acquisition costs, such as, one company may value customer re-orders whereas one may consider a customer’s ethics and backgrounds as high value.

2.2 Proposed Solution

The solution proposed by us makes use of a very diverse data set consisting data of about 60,000 customers in the form of what kind of products do they like to purchase, what are their educational backgrounds, their income, their marital status, and many stores related parameters such as the size of the store, the types of products available in a store, etc. With the help of this balanced dataset, without any null values, and Machine Learning Predicting Algorithms, we can accurately predict the estimated cost of acquiring a new customer.

THEORETICAL ANALYSIS

3.1 Block Diagram


Data Collection
Pre-Processing
Model Preparation and Evaluation
Deploying the Model
Predicting Customer Acquisition Cost based on user inputs

3.2 Hardware / Software Design

Technologies used in this project:

- exploratory analysis, data pre-processing, - python, numpy, pandas, seaborn, matplotlib, missingno
- model development and evaluation – linear regressor, random forest regressor, LASSO regressor
- model deployment – python, streamlit

EXPERIMENTAL INVESTIGATIONS

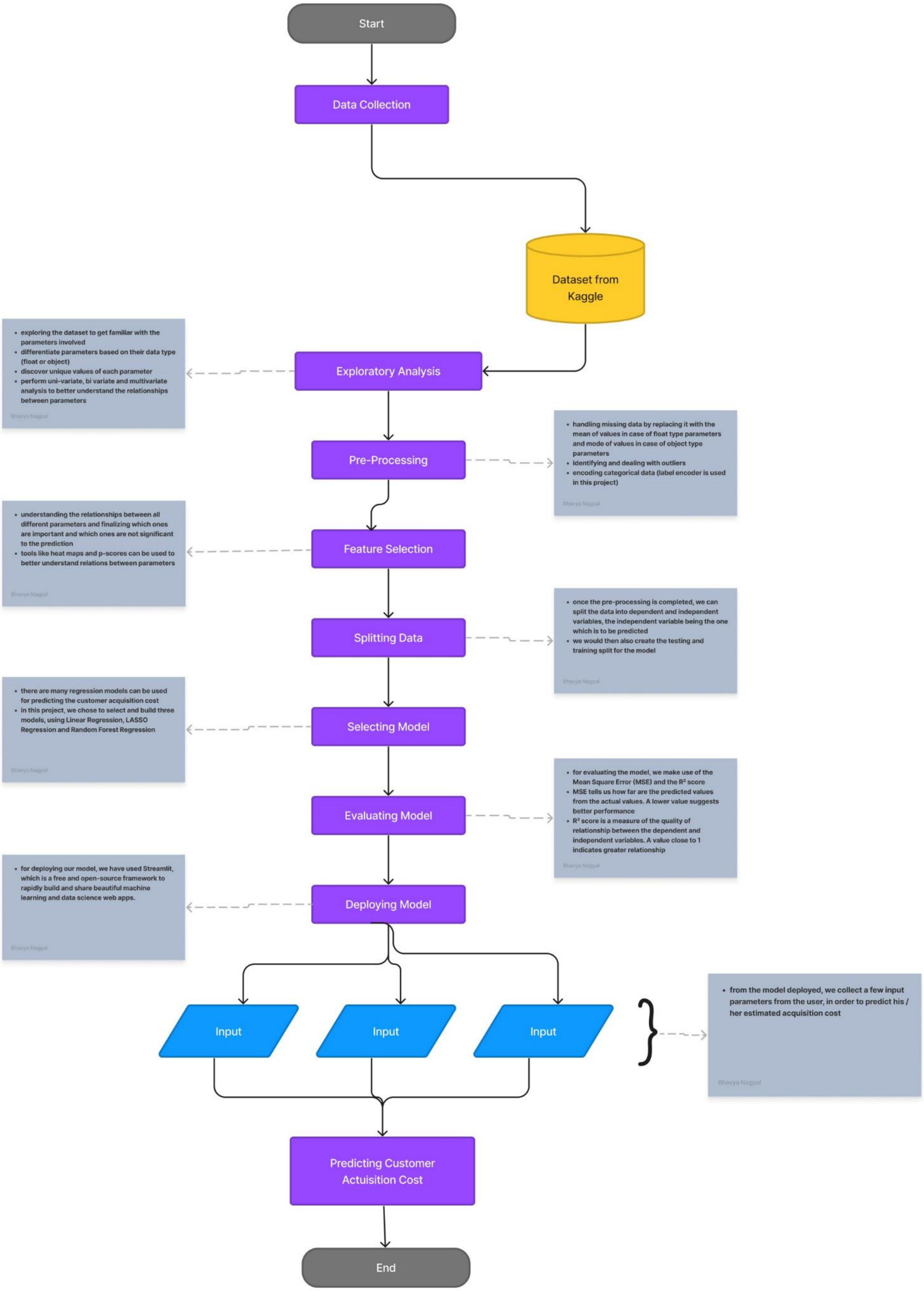
While exploring and understanding the dataset, we discovered that the dataset covers ten stores across America, three of which are in the United States, six in Mexico and one in Canada. The dataset consists of a total of 19 cities, ten from the US, seven from Mexico and two from Canada. USA sold the greatest number of products, followed by Mexico and lastly by Canada.

While analysing existing customers, we found out that the number of male and female customers were nearly equal. The dataset was also evenly split between married and unmarried individuals. The dataset contained about 35,000 customers who owned a home whereas about 25,000 who did not, and most of the customers had a high school degree, while a little over 2,500 were graduates.

While analysing the type of products sold, we saw that there was a huge sale of food products, as opposed to non-consumables and drinks.

After the necessary pre-processing and removing the outliers, we found out that the most factors that affect the cost of acquiring a customer include promotion name, media type, and store features like grocery sqft. The size of the store and the different sections it contained also proved to be extremely valuable.

FLOWCHART



- exploring the dataset to get familiar with the parameters involved
- differentiate parameters based on their data type (float or object)
- discover unique values of each parameter
- perform uni-variate, bi variate and multivariate analysis to better understand the relationships between parameters

- handling missing data by replacing it with the mean of values in case of float type parameters and mode of values in case of object type parameters
- identifying and dealing with outliers
- encoding categorical data (label encoder is used in this project)

- understanding the relationships between all different parameters and finalizing which ones are important and which ones are not significant to the prediction
- tools like heat maps and p-scores can be used to better understand relations between parameters

- once the pre-processing is completed, we can split the data into dependent and independent variables, the independent variable being the one which is to be predicted
- we would then also create the testing and training split for the model

- there are many regression models can be used for predicting the customer acquisition cost
- in this project, we chose to select and build three models, using Linear Regression, LASSO Regression and Random Forest Regression

- for evaluating the model, we make use of the Mean Square Error (MSE) and the R^2 score
- MSE tells us how far are the predicted values from the actual values. A lower value suggests better performance
- R^2 score is a measure of the quality of relationship between the dependent and independent variables. A value close to 1 indicates greater relationship

- for deploying our model, we have used Streamlit, which is a free and open-source framework to rapidly build and share beautiful machine learning and data science web apps.

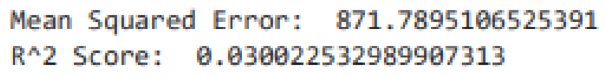
- from the model deployed, we collect a few input parameters from the user, in order to predict his / her estimated acquisition cost

RESULT

After pre-processing the dataset and splitting the data into dependent and independent variables, we create the training (80%) and testing (20%) split for our model. Since we would be predicting the customer acquisition costs, regression models would be the choice of algorithm. We implemented three regression models, namely, Linear Regression, LASSO Regression and Random Forest Regression.

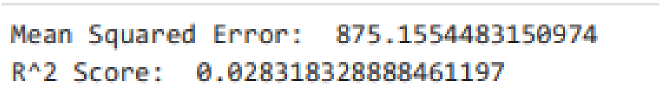
For evaluating our models, we used Mean Square Error (MSE) and R^2 Scores. The results given by each of our models are displayed below:

Linear Regression:



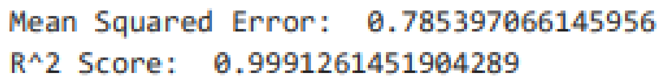
```
Mean Squared Error:  871.7895106525391
R^2 Score:  0.030022532989907313
```

LASSO Regression:



```
Mean Squared Error:  875.1554483150974
R^2 Score:  0.028318328888461197
```

Random Forest Regression:



```
Mean Squared Error:  0.785397066145956
R^2 Score:  0.9991261451904289
```

We can see that Random Forest Regression worked the best for our dataset, with an MSE score of 0.7 and R^2 score of 0.99, indicating that the independent variable can strongly identify and predict the dependent variables.

ADVANTAGES AND DISADVANTAGES

Predicting customer acquisition costs can offer a great deal of assistance to companies trying to expand their reach towards customers. It helps them allocate their budget effectively by enabling them to estimate their expenses early on. It also helps them identify areas of strength and weaknesses in their marketing strategies, which in turn would allow them to quicky improve their approach towards acquiring more customers, without a great loss of time or finances.

While predicting costs can be extremely helpful, a huge part of it greatly relies on the availability and quality of the data. Inaccurate date can impact the accuracy of predictions and affect the overall effectiveness of the strategies. Besides, consumer behaviour and the competition in the market changes rapidly, making it difficult to accurately predict customer acquisition cost over time.

APPLICATIONS

The insights gained through this project can prove to be helpful for the Food Mart of USA in order to make more informed and impacting decisions, backed by pre-existing data, in order to reach out to a greater audience. Apart from gaining customers, the company can optimise their spendings on promotions and advertisements to effectively acquire new customers.

CONCLUSION

This project lets us predict the cost of acquiring customers in Food Marts of the USA using the data of 60,000 customers. Our goal was to understand how multiple factors like income, store features, type of product sold, etc would affect the cost of acquiring a customer.

After thoroughly analysing the data, we found out that some of the most important factors that affect the cost include promotion name, media type, and store features like grocery sqft. After applying multiple regression techniques, we obtained positive results using the Random Forest Regressor, which gave us an R^2 score of 0.999, indicating a strong correlation between the independent and dependent variables.

Predicting customer acquisition costs using machine learning techniques can be used as an effective tool in conjunction with other business strategies and research, to make well-informed decisions.

FUTURE SCOPE

The future scope of the research paper involves several promising avenues. Future research can explore advanced machine learning techniques, such as deep learning and ensemble methods, to enhance the accuracy and predictive power of models. Additionally, investigating industry-specific analysis can provide insights into variations and factors influencing customer acquisition costs within different contexts. Incorporating dynamic factors and time-dependent variables can capture temporal variations, while feature engineering and selection can identify relevant variables impacting acquisition costs. The research can also focus on translating predictive insights into actionable recommendations for marketing managers, developing decision support systems, and conducting comparative analyses of different machine learning algorithms and models. By pursuing these directions, the paper can contribute to improved decision-making and marketing strategies in the context of customer acquisition.

BIBLIOGRAPHY

[1] Schröder, B. (2023). New Customer Acquisition and Lasting Customer Loyalty Through Holistic Branch Design. In Multisensory in Stationary Retail: Principles and Practice of Customer-Cantered Store Design (pp. 247-259). Wiesbaden: Springer Fachmedien Wiesbaden.

[2] Camacho-Vallejo, J. F., Dávila, D., & Nucamendi-Guillén, S. (2023). A hierarchized green supply chain with customer selection, routing, and nearshoring. Computers & Industrial Engineering, 178, 109151.

[3] Calder, B., Malthouse, E., & Omatoi, J. (2023). Reconciling the Customer Equity and the Brand Equity Perspectives on the Financial Value of Marketing. Available at SSRN 4321991.

[4] Cherchye, L., De Rock, B., Dierynck, B., Kerstens, P. J., & Roodhooft, F. (2023). A DEA-based approach to customer value analysis. European Journal of Operational Research.

[5] Harsh, I. S., Sharma, A., Sharma, C., & Garg, R. (2023). Intelligent Neurological Based Email Categorization Using SVM to Increase Customer Support Efficiency. Journal of Pharmaceutical Negative Results, 4166-4171.

[6] Saran Kumar A., Chandrakala D., A Survey on Customer Churn Prediction using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887) Volume 154 – No.10, November 2016

APPENDIX

Source Code

Github link:

https://github.com/Gesskay/CAC_Predictor_Team265/tree/master