Project Cost prediction Final Report

Mr. Anwesh Reddy
Mentor

Piyush Srivastava
IIIT Delhi Student
piyush.srivastava906@gmail.com

Utkarsh Srivastava
IIIT Delhi Student
utkarsh.srivastava943@gmail.com

Byrapu Shirisha
IIIT Delhi Student
bsc.chowdary@gmail.com

Abstract: Companies today need to find a balance between the costs of Media campaigns, and the point where it has reached its target goal. And also, not to just do a post analysis of what happened, but to actually predict how much cost the Organization will incur based on its target. while this was extremely difficult to do earlier and it's still challenging today, technology can help in this.
With the advent to ML, and fact-based decision-making tools, that can compute and predict targets based on multiple variables, companies are now finding this technology as an investment, even though this technology has a cost of its own to implement.

The challenges in implementing the prediction of correct costs are many, of which include large amounts of data collection, maintaining the data, and validation of the correctness of the data. But these supporting functions are also now provided by technology, making the resulting ML implementation a real possibility.
Machine Learning helps retailers to predict the future through simulating scenarios that predetermine the outcomes and identify the crucial action areas. In Cost prediction on acquiring customers, we need to predict the cost of media campaigns in food mart of the USA.

Base model selected for this research is Linear Regression model, Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine

Keywords: Cost Prediction, Customer Acquisition, Regression Analysis, Campaign Data, Machine Learning Models, Hyperparameter optimization

1. Introduction

CRM : In recent days CRM has become much popular and must use technology for any size of businesses, business are quickly noticing CRM incorporated with Machine Learning is a great combination which offers incredible feature, CRM helps business to capture and maintain leads, prospect and customer data, CRM included with ML will help business to do predictive analysis which helps to identify and segment the data based on various attributes like geographic area, family size, income, transactional history etc., this analysis will be used for better campaigning techniques which leads to better sales and profits
Customer Acquisition is process of getting potential customers to buy our products. A strong customer acquiring strategy is to attract leads through different channels, nurture them and convert them into prospects and finally convert them into customers by making them purchase products. One of the first and foremost things is to reduce cost to increase profitability of the business, once such cost is Acquisition cost.
Customer Acquisition Cost (CAC) : CAC is the cost related to acquiring new customers, it is calculated based on cost spent to get new customers divided by numbers of customers acquired and their sales. If we do segmentation of customers acquired by various channels and review sales processed for those customers during that period, we will get to know what is the amount spend on those customers and what is the revenue made upon them, this helps to where we are overspending or underspending cost on acquiring new customers

2. Literature review

Keswani, K[1] has provided his analysis on today's data driven world, how to leverage the information you have about the customer to understand what they want approach them with the right campaign based on their preferences, reduce the churn rate and customer acquisition cost to enhance the customer experience and improve the overall Profitability of the business.

Malmberg [2] proposed Random Forest and XGBoost algorithms based on his previous studies, he experimented above 2 algorithms by an iterative process which began with data processing followed by feature selection, training of model and testing the model, as per final result Random Forest proved best in terms of accuracy.

Dixit[3] as per their research analysis, they proposed K-means clustering machine learning algorithm for customer segmentation, customer segmentation is performed on the company's customers data and with the help of K-means clustering machine learning algorithm, customers are divided using features like total spending and annual income, this study also proves that the dividing customers on the basis of behavioral characteristics is a better solution for existing customer segmentation problem and K-means clustering algorithm is identified as a good choice for this approach.

Dutta.H[4] proposed a prediction model for customer acquisition, as per their research Unsupervised Learning(K-Means clustering) has been used for partitioning the dataset into K clusters and then apply supervised learning model for predicting whether a person would respond to marketing event or not, as per their report xgboost classifier performed much better than gradient boosting classifier after good feature engineering (e.g. identifying categorical variables for one-hot encoding)

Balaji et al [5] describes the use of different algorithms in conjunction, and compares them. Like various models ANN, Linear Regression, KNN, used for Business Intelligence prediction. Sarker at al [6] describes the various ML techniques that are used for solving real world problems that have Structured/ Unstructured data sets, with Multiclass classification, with approaches like Naive Bayes (NB), K-nearest neighbors (KNN), Regression Analysis using multiple linear regression, Clustering using K-means clustering.

3. Materials and Methods

3.1 Predicting media cost methods and techniques: Sheng Yu and Subhash Kak [7] used the Regression method and K-nearest neighbor classifier for predictions. Mahdieh et al [8] describes KNN models that can be used to predict prices based on various factors, including socio-economic factors. The KNN method uses attributes similarity to predict the values of any new data points. It checks how alike is a data point (a vector of features) in the test data set from other available data points in the training data set. Mahdieh states KNN method is an efficient and accurate technique that can be used in a variety of applications such as finance, economic forecasting or prediction.

Sebastião [9] describes forecasting of dependent variables (like value) that can be done with ML under changing market conditions. Sebastião has used Linear Multivariate Regression Models along with others to produce forecasts of the dependent variable

3.2 Data Mining Techniques: Data mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich database of history of their customer transactions for millions of customers dating back a number of years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.
The classic examples of applications of data mining are the influenza trend forecast service pushed out by Google and "election of big data" [10] of the Obama team. Other Domestic scholars have also started related research. For example, Meng Xiaofeng who systematizes and concludes the concepts, technologies and challenges of big data management; Hou Jingchuan who studies the quotation of data in the age of big data and has a deeper analysis and discussion on its current situation, latest development and future improvement. Today commonly used algorithms of data mining can be divided into several kinds of classification, cluster, association rules and time series prediction. Data mining is used in aspects of banking service, telecommunication, information security and scientific study. Furthermore, the popular data mining tools are Weka, statistical analysis software Spass, Rapidminer, Knime, Keel, Clementine, Orange, Tanagra and so on. Sorting algorithms that are often used in data mining are mainly *Linear regression algorithm, Logistic regression algorithm, Bayesian decision theory and*

classifier, Support Vector Machine proposed by Cortes and Vupnik in 1995.

Clustering algorithms commonly used in data mining are mainly hierarchical clustering algorithm, partition clustering algorithm, clustering algorithm based on density, clustering algorithm based on network, clustering algorithm based on model which may also include statistical method and neural network method.

3.3 Linear Regression for predicting media cost: The algorithm uses linear regression for prediction and uses the Akaike criterion to select models; the algorithm could work with weighted instances. This method of regression is simple and provides an adequate and interpretable description of how the input affects the output. It models a variable Y (a response value) as a linear function of another variable X (called a predictor variable); Given n samples or data points of the form (x1, y1), (x2, y2), ..., (xn, yn), where xi ∈X and yi ∈Y, predictive regression can be expressed as Y = α + β X, where α and β are regression coefficients. Assuming that the variance of Y is a constant, the coefficients can be solved using the least squares method. This minimizes the error between the actual data point and the regression line.

$$\beta = \frac{\left[\sum (x_i - mean_x)(y_i - mean_y)\right]}{\left[\sum (x_i - mean_x)^2\right]} \quad \text{and } \alpha = mean_y - \beta * mean_x$$

where mean of x and mean of y are the mean values for random variables X and Y given in a data training set. The X variable is the input value (independent) and Y is the response output value (dependent) that depends on X.

3.4 Applications of media cost prediction: Predictive analytics uses data models, statistics, and machine learning to predict future events. In marketing, this can be used to make better decisions regarding media planning and buying. the following scenarios where we use media cost prediction:

3.5 Churn Prevention:
[11] It proves to be expensive as the cost of acquiring a new customer is much higher than retaining the existing customer.
These models help prevent churn in your customer base by analyzing the

3.6 Customer Lifetime Value:
[12] The great promise of mass media campaigns lies in their ability to disseminate well defined behaviorally focused messages to large audiences repeatedly, over time, in an incidental manner, and at a low cost per head. (1)

It is pretty challenging to identify the customer in the market who is most likely to spend large amounts of money consistently over a long period.

This kind of data through predictive analytics use case allows the business to optimize their marketing strategies to gain customers with the most significant lifetime value towards your company and product.

3.7 Customer Segmentation:[13] Customer segmentation enables you to group the customer by shared traits. Profound use of predictive analytics techniques helps target the markets based on accurate insights and indicators and analyze the segments of those most interested in what your company offers. Using these predictive analytics applications, you can make data-driven decisions for each part of your business. The same data also enables you to potentially identify the entire markets that you didn't even know existed.

[14] Determining your primary marketing goals and customers is a critical use case for predictive analytics. It only provides an incomplete picture of what your marketing approach should be.

4. IMPLEMENTATION:

4.1. Dataset, data collection, data preprocessing:

Food mart X is a chain of convenience stores in the United States. The private company's headquarters are in Mentor,Ohio, and currently, approximately 325 stores are in the US. Convenient food mart operates on the franchise system.
The dataset we have chosen is "PREDICT COST ON MEDIA CAMPAIGNS IN FOOD MART OF USA". Source of data for this research is taken from Kaggle website. The dataset can be downloaded by the link below

https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart

4.2. DataSet Summary:

The data downloaded consists of 3 files i.e **Train**, **Test** and **data dictionary** respectively. In the dataset, there are **36256** samples in the "**Train.csv**" and **12086** samples in the "**Test.csv**".
There are 17 Categorical and 23 Continuous attributes

4.3. EDA:

Based on the heat map plot, identified continuous variables which have strong correlation with each other and removed them from the data set to avoid redundant data. We choose the thresholds with 0.8 and -0.8, so features that greater than 0.8 or lesser than -0.8 are considered as strong positive correlation and strong negative correlation respectively. Following columns were removed from data set, "salad_bar"," meat_sqft","gross_weight","store_sales", "store_cost", "store_sqft".
Categorical variables were encoded based on One Hot Encoding and Label Encoding techniques.

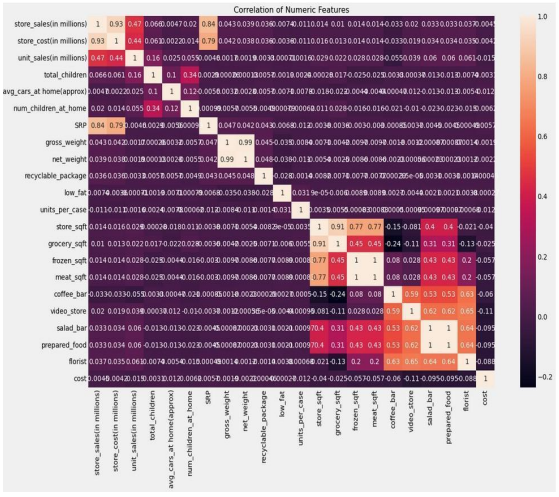


Fig. 1

5. Data Preparation for model building:

Segregating the target column i.e., cost in "Y" and the rest of remaining columns into "X". Then we call the **"train_test_split"** function and divide our dataset into 2 parts, 70% training data and 30% testing data. After that we will be normalizing the data using *Feature Scaling (Standard Scaler)*, so that all the data should be at the same level.

6. K Best Features:

We used Linear Regression estimator, features with the highest absolute "coef_" value as shown in below fig2, features selected from select model are [Store_type,store_city,video_store,product attachment, cash register handout, Sunday paper, TV, radio, street handout]

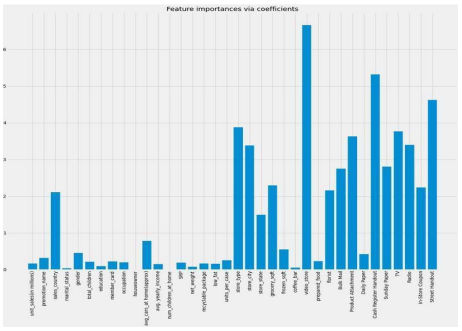
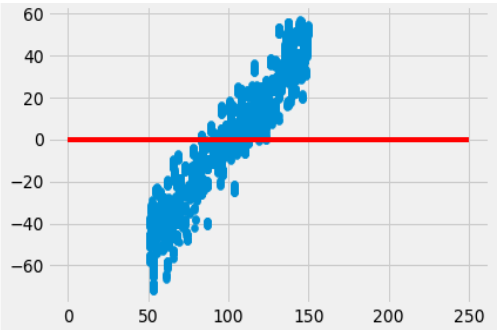


Fig: 2

7. **Linear Regression:** Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

   By Plotting the scatter graph we can clearly see the linear relation between **x="y_test", y="y_test-y_pred_test"**



8. **Decision Tree Regressor:-**[15]A decision tree is a data structure that is built using nodes (containing values or conditions) and edges (connecting all nodes). The decision tree is constructed based on a dataset containing attributes or features which classify the raw data for each record. Each node in the tree can be either a decision node that makes decisions or a leaf node that gives the outcome
9. **Random Forest:**[16] Random Forest is a data mining algorithm that uses decision trees to draw conclusions based on randomly sampled data. This model can also make predictions using regression to classify

vast amounts of data, while a variety of transformations may be applied to hold onto samples that fall outside the range expected

10. **XGBoost Regression:** The prediction performance of the XGBoost method is evaluated by comparing observed and predicted PM2.5 concentration using three measures of forecast accuracy. The XGBoost method is also compared with the random forest algorithm, multiple linear regression, decision tree regression and support vector machines for regression models using computational results. The results demonstrate that the XGBoost algorithm outperforms other data mining methods.

11. **KNN (K-Nearest Neighbor):** [17]K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well-suited category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

12. **Results**: We have evaluated data with a couple of different models too based on parameters mentioned below and the table shows the results of each module performed and Train and Test Data.

Model	R2 Score	MAPE	MSE	RMSE	MAE
Linear Regression	0.097000305618 43187	0.284101336483 41507	817.8603343304433	4.91860651664392 1	24.1926900655720 44
Lasso Regression	0.096966262365 02959	0.284163452870 5908	817.8911678144217	4.91920913029260 9	24.1986184675541 66
Ridge Regression	0.097000298559 52941	0.284101346805 341777	817.8603407237969	4.91860661863671 1	24.1926910688968 63
ElsaticNet Regression	0.096830702711 87555	0.284101346805 341777	817.8603407237969	4.91860661863671 1	24.1926910688968 63
Decision Tree	0.998522815388 9094	0.000262661446 3205253	1.337908204633205	0.18956870373673051	0.035936293436424 31
Random Forest	0.998897519257 417	0.000644132407 1288373	0.998533304420848 3	0.28194379860356245	0.079492305571006 17
XgBoost	0.907782695926 7506	0.000644132407 1288373	0.998533304420848 3	0.28194379860356245	0.079492305571006 17
KNN	0.491669848131 779 05	0.194935278226 024 88	460.402224434 7664	4.083704154847434	16.676639624318195

MAPE: Mean Absolute Percentage Error
MSE: Mean Squared Error
RMSE: Root Mean Squared Error
MAE: Mean Absolute Error

13. Design, develop and train the Pipeline:

1. **Creation of ML pipeline**
Make pipeline ()" function that will create the pipeline from sklearn pipeline import Pipeline, make pipeline
2. **Random Forest Base Model as was best model use as reference**
3. Filtered sample from train data as test data is already processed.
Typically, when our code run to data is not processed.
4. Creating activities to put in the pipeline used (Function transformers)
5. **Creating pipeline using the activities.**
6. Call the fit () function on the pipeline object
7. Saving pipeline for deployment to production

Pipeline:

1. We need to tie together many different processes that we use to prepare data for machine learning based model
2. It is paramount that the stage of transformation of data represented by these processes are standardized
3. Pipeline, along with the Grid search CV helps search over the hyperparameter space applicable at each stage.

Hyper parameters are like handles available to the modeler to control the behavior of the algorithm used for modeling
Hyper parameters are supplied as arguments to the model algorithms while initializing them.

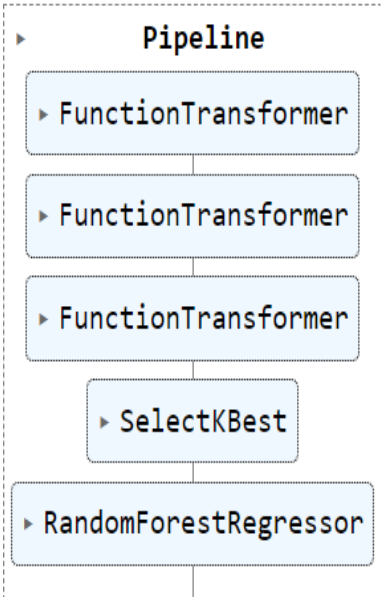
Linear Regression Base Model
Train R2 Score of Linear Regression: 0.0954127994241738 Test R2 Score Linear Regression: 0.09700030561843187

Random Forest Base Model
Train R2 Score of Random Forest: 0.9998232139847923 Test R2 Score Random Forest: 0.998913645321768

Random Forest Model score after hyperparameter tuning

Best R2 score: .9971852037905068 Best Hyper Parameters: 'bootstrap': True, 'max_depth': 17, 'estimator's': 30}

Train R2 Score of Random Forest:
0.9998232139847923 Test R2 Score Random Forest: 0.998913645321768



14. BASED ON THE CUNDUCTED STUDY, KEY FINDINGS ARE:

1. There is considerable improvement in the r2 score from the base regression model when random forest model is used.

2. Best performance w.r.t r-square score is obtained using random forest model.

3. There is no considerable impact on performance after Hyperparameter tuning in the random forest model

15. FUTURE IMPROVEMENTS: -

1. As a next step, we can explore more features which can potentially impact the media cost.

2. Further we can try applying PCA (principal component analysis) to reduce multicollinearity.

3. Also, we can explore state of art deep neural network to capture variations in the data.

4. **Conclusion:** Based on results Random Forest and Decision tree regressor are identified as best performing models with 99% accuracy

References:

[1]Keswani, K. (2022, October 10). *Customer analytics in the data dominating world!* Times of India. https://timesofindia.indiatimes.com/blogs/relatable-technology-club/customer-analytics-in-the-data-dominating-world/

[2]Malmberg, O. (n.d.). *Using machine learning to detect customer acquisition opportunities and evaluating the required organizational prerequisites*. Diva-portal.org. Retrieved October 16, 2022, from https://www.diva-portal.org/smash/get/diva2:1366207/FULLTEXT01.pdf

[3]Dixit, S. (n.d.). *Customer Segmentation using machine learning*. Researchgate.net. Retrieved October 20, 2022, from Dixit, S. (n.d.). Researchgate.net. Retrieved October 20, 2022, from https://www.researchgate.net/publication/356756320_Customer_Segmentation_Using_Machine_Learning

[4]Dutta, H. (2020, September 17). *Creating a prediction model for customer acquisition*. The Startup. https://medium.com/swlh/creating-a-prediction-model-for-customer-acquisition-3517f538ef66

[5] Balaji T.K., Chandra Sekhara Rao Annavarapu, Anneshree Bablani, Machine Learning Algorithms for Social Media Analysis [March 25, 2021]

[6] Iqbal H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions [March 2021]

[7] Sheng Yu and Subhash Kak, A Survey of Prediction Using social media[March 2012]

[8] Mahdieh Yazdani, Machine Learning, and Hedonic Methods for Real Estate Price Prediction [October 15, 2021]

[9] Helder Sebastião, Forecasting and trading cryptocurrencies with machine learning under changing market conditions, (2021) 7:3

[10]. M. Scherer. Inside the Secret World of the Data Crunchers Who Helped O bama Win. [EB/ O L]. (2012- 11-07)[2013- 03- 06]. http://swampland.time.com/ 2012/ 11/ 07/inside-the-secret-world-of-quants-and-data-crunchers- who-helped- obama- win/

[11] Carol Anne Hargreaves (2019) - A Machine Learning Algorithm for Churn Reduction & Revenue Maximization

[12] Srikanta Patnaik, Xin-She Yang,Ishwar K. Sethi (2019) - Advances in Machine Learning and Computational Intelligence

[13]Sunčica Rogić and Ljiljana Kašćelan - Class Balancing in Customer Segments Classification Using Support Vector Machine Rule Extraction and Ensemble Learning

[14] Lancet. 2010 Oct 9; 376(9748): 1261–1271

[15] European Journal of Molecular & Clinical Medicine

[16]C.H. Bryan Liu 1 , Benjamin Paul Generalizing Random Forest Parameter Optimisation to Include Stability and Cost [July 2017]

[17]J.D. McCaffrey, Restricted boltzmann machines are difficult to explain, Software Research, Development, Testing, and Education,