*Advances in Parallel Computing Algorithms, Tools and Paradigms*
*D.J. Hemanth et al. (Eds.)*

*© 2022 The authors and IOS Press.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/APC220030*

# Prediction of Likely Customers for Car Industries Using K-Means Clustering Compared with Logistic Regression

V.UpendraReddy[a] and Dr.S.John Justin Thangaraj[b,1]
[a]*Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Professor, Department of CSE, Saveetha School of Engineering,*
[a,b]*Saveetha Institute of Medical and Technical Sciences,*
[a,b]*Saveetha University, Chennai, Tamilnadu. India*

**Abstract.** This study represents a customer segmentation model which helps them to group the customers for the car industry with the same market characteristics. The study contains 2 groups i.e, the Logistic Regression model is developed in the first group and the K Means clustering model, an unsupervised machine learning algorithm is developed in the second group. Each group has a sample size of 200 and the study parameters include alpha value0.05, beta value 0.2, and the power value 0.8. The accuracies of each model are compared with others for different sample sizes. This article is an attempt to improve the accuracy of customer segmentation using the K Means clustering, an unsupervised clustering machine learning algorithm. The proposed model has improved accuracy of 87.4% with $p < 0.05$ in segmenting customers than the existing model of 85%. This innovative prediction model helps to know the future customers and their needs and take innovative decisions to fulfill them. The outcomes of the proposed model are compared with the Logistic Regression algorithm and the proposed model confirms to have higher accuracy than the Logistic Regression algorithm.

**Keywords.** Customer Segmentation, Innovative prediction model, K Means Clustering, Logistic Regression, Unsupervised machine learning, Car industry.

## 1. Introduction

Nowadays, it is very important for an industry to know its product's performance and customer satisfaction[1]. It helps to improve the goodwill and sales of the company[2]. Customer satisfaction plays a very important role between the effects of Customer Relationship management and customer faith towards the industry[3]. A prediction model is developed to identify the most likely customers for the car industry using K Means Clustering and it is the existing models are not accurate in clustering customer data of large size. The existing experience of my team is in machine learning algorithms, python, and data mining[4]. Using this experience, an innovative prediction model is proposed to improve the accuracy of identifying unique customers.

[1]Dr.S.John Justin Thangaraj, Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai, India. Email:johnjustinthangarajs.sse@saveetha.com

## 2. Materials and Methods

The study was carried out at Saveetha School of Engineering. This study consists of two sample groups i.e, Logistic Regression and K Means Clustering. Each group contains 200 samples with a pretest power of 0.8. The sample size was collected by using previous results from[5] in clinicalc.com by keeping threshold 0.05, G power of 80%, confidence interval at 95%, and enrolment ratio as 1.

The dataset used for classification is taken from the car company through Kaggle©[6], an open-source data repository[7]. Table 1 represents a little preview of the dataset. It contains socio-economic details of the car company's previous customers. The dataset contains 6 columns. In this study, Jupyter notebook is used for python execution. A correlation matrix graph is plotted to find the important features[8]. The dataset is checked for missing and not a number (NAN) values and will be removed if any are present. The required libraries are imported.

**Table 1**: Car customer dataset sample collected from Kaggle©

| ID | Product_Name | Age | Gender_Desc | Price | cibil_score |
|---|---|---|---|---|---|
| 1 | CHEVROLET SPARK | 41 | Male | 237342 | 594 |
| 2 | FORTUNER 4 WD | 42 | Male | 2400000 | 754 |
| 3 | TOYOTA - FORTUNER | 40 | Male | 2200000 | 779 |
| 4 | NISSAN MICRA | 35 | Male | 365750 | 818 |
| 5 | XYLO E8 | 34 | Male | 548750 | 853 |
| 6 | FORD FIGO | 33 | Male | 329250 | 777 |
| 7 | MARUTI SWIFT | 45 | Male | 482750 | 817 |
| 8 | MARUTI RITZ | 46 | Female | 461500 | 789 |
| 9 | HYUNDAI VERNA | 34 | Male | 437250 | 828 |

### *2.1. K Means Clustering*

K-Means clustering is an unsupervised learning algorithm that divides the dataset into a certain number of non-overlapping clusters which have similar properties. Each cluster is associated with a centroid.

The step by step process for k means clustering is as follows: Load the dataset and remove missing and not a number (NAN) values. Then the Elbow method is used to find the optimal number of clusters. It applies k-means clustering on the dataset for k values which range from 1 to 10[9]. The Sum of Squared Error (SSE) is calculated for each value of k. A graph is plotted using SSE against k values. A sharp turning point that looks like an elbow in the plot is considered as the optimal number of clusters. Now, give the k value in k means class and fit the dataset into it for clustering. Initialize the centroids randomly and group each data point to its nearest centroid[10]. Then, continuously calculate the mean values of all data points of a cluster and move the centroids. Iterate the model until there is no change in centroid positions[11]. This innovative model returns the cluster number for each sample and the clusters are displayed. Then accuracy is checked for the clustering.

### *2.2. Logistic Regression*

Logistic Regression takes a set of independent variables and predicts the target variable in which the output will be 0 or 1[12]. A sigmoid function which is also called a logistic function is used. It gives the probability of the dependent variable. It returns 0 if the probability is less than 0.5 and 1 if the probability is above 0.5[13]. Regularization helps to avoid overfitting of parameters to improve the prediction power of the model.

The step by step process of the Logistic Regression algorithm is as follows:

- The independent variables in this study are gender, income, age, kilometers driven, price, and the dependent variable is purchased.
- Get the dependent and independent variables from the dataset. Then, Split the dataset into train and test sets. Scale the sets using the standard scaler for improved training of the model.
- Import LogisticRegression class from sklearn library and fit the training set into the model for training it. Now, fit the test set to the model to get predictions.
- Then, check the accuracy of the predictions.

### *2.3. Test Procedure*

For training the Logistic Regression, the test set size is about 20% of the total dataset and the remaining 80% is used for the training set. The whole dataset is fitting for training the K Means Clustering model. Accuracies of both models are tested with different sample sizes from 50 to 1000 [14].

### *2.4. Statistical Analysis*

SPSS Version 26 software tool was used for statistical analysis. The independent sample t test was performed to find the mean, standard deviation and the standard error mean statistical significance between the groups. Cost is the dependent variable and Age, Gender_Desc, Owners, Income, and Driven_Kms are the independent variables in the dataset. Cost is calculated using independent variables in Logistic Regression and Cost, Age, Gender_Desc, Income are used to get the respective cluster of customers in NKMC. Standard deviation, standard mean errors were calculated using the SPSS Software tool[15].

## 3. Result

The group statistical analysis on the two groups shows that K Means Clustering has more mean accuracy than the other and its standard error mean is slightly less than Logistic Regression. In the independent sample test, the significance of both algorithms when the equal variance is assumed is 0.025. Figure 1 represents the bar chart of accuracies with standard deviation error is plotted for both the algorithms. The K Means Clustering algorithm scored an accuracy of 87.4% and Logistic Regression has scored 85%. Table 2 & Table 3 represent the independent sample test for K Means Clustering and Logistic Regression respectively.

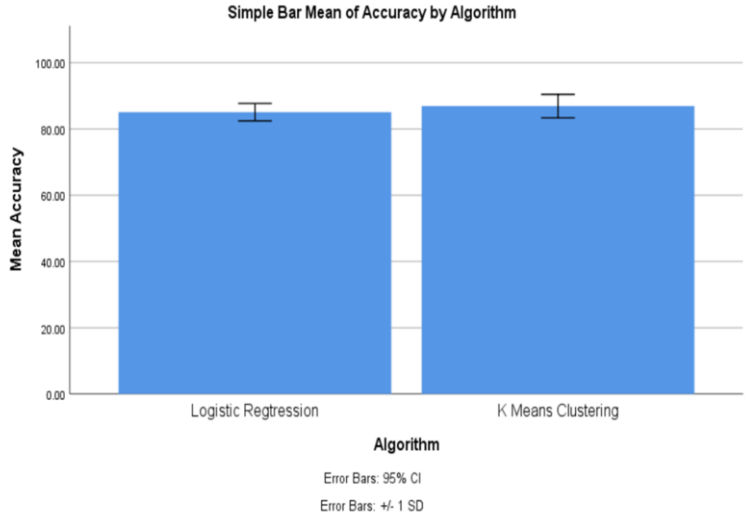


**Figure 1.** Bar chart with K Means Clustering algorithm with accuracy of 87.4% and Logistic Regression has scored 85%.

**Table 2**. Group Statistic analysis, representing Logistic Regression (mean accuracy 85.0750%, standard deviation 2.62114) and K Means clustering (mean accuracy 87.4850% standard deviation 3.54450)

| | Algorithm | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Accuracy | Logistic Regression | 20 | 85.0750 | 2.62114 | 0.58611 |
| | K Means Clustering | 20 | 87.4850 | 3.54450 | 0.79257 |

## 4. Discussion

From the results of this study, K Means Clustering is having better accuracy than the Logistic Regression algorithm. K Means Clustering has an accuracy of 87.4% whereas Logistic Regression has an accuracy of 85%. Most of the researchers used Random Forest to identify customer churns[16]. The electricity load curve of electricity customers is clustered by Pan using R - based parallelized K Means algorithm[17]. Sokol proposed a model for segmenting customers alongside the traditional methods[18].

Chan proposed a novel approach for the clustering of consumption behavior[19]. All the previous literature supports the proposed model and its results are better than theirs. The factors that may affect the accuracy of K Means Clustering are limited data set for analysis, variation in the amount of training dataset, and inclusion of more relevant attributes. The availability of more datasets related to branded cars, effective data preprocessing techniques, and the combination of K Means Clustering with other machine learning algorithms like a decision tree and artificial neural networks may give further accurate results in the identification of unique customers.

**Table 3**. Independent Sample Test for Logistic Regression and K Means Clustering (mean difference -1.8100 and standard deviation error difference 0.98574 with significance 2-tailed 0.074 and 0.075 respectively).

| Accuracy | Equality of Variances | | T-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig. (2-tail) | Mean Difference | Std. Error Difference | 95% Conf. Interval Lower | 95% Conf. Interval Upper |
| Equal Variances assumed | 2.230 | 0.025 | -1.836 | 38 | 0.074 | -1.8100 | 0.9857 | -3.805 | 0.1855 |
| Equal Variances not assumed | | | -1.836 | 34.9 | 0.075 | -1.8100 | 0.9857 | -3.811 | 0.1911 |

## 5. Conclusion

The K-Means clustering has been proved to predict the most likely customers of the car company more significantly than Logistic Regression. It can be used in any car business to group customers and find the ones who are more interested to buy the company products.

## References


[1] Why measure customer satisfaction? Customer Satisfaction Measurement for ISO 9000: 2000 2007; 12–19.

[2] Liu S, Wang X, Collins C, et al. Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE Trans Vis Comput Graph 2019; 25: 2482–2504.

[3] Sulaiman, Musnadi S. Customer Relationship Management, Customer Satisfaction and Its Impact on Customer Loyalty. Proceedings of the 7th International Conference on Multidisciplinary Research. Epub ahead of print 2018. DOI: 10.5220/0008892606920698.

[4] Irene DS, Shiny Irene D, Surya V, et al. An Intellectual Methodology for Secure Health Record Mining and Risk Forecasting Using Clustering and Graph-Based Classification. Journal of Circuits, Systems and Computers 2020; 2150135.

[5] Wang Z, Zuo Y, Li T, et al. Analysis of Customer Segmentation Based on Broad Learning System. 2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). Epub ahead of print 2019. DOI: 10.1109/spac49953.2019.237870.

[6] Birla N. Vehicle dataset, https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho (accessed 6 July 2021).

[7] Aman Miglani, https://www.kaggle.com/datatattle/home (accessed 8 July 2021).
[8] Iman RL, Davenport JM. An iterative algorithm to produce a positive definite correlation matrix from an approximate correlation matrix (with a program user's guide). Epub ahead of print 1982. DOI: 10.2172/5152227.
[9] Borgelt C, Yarikova O. Initializing k-means Clustering. Proceedings of the 9th International Conference on Data Science, Technology and Applications. Epub ahead of print 2020. DOI: 10.5220/0009872702600267.
[10] Wu J. K-means Based Consensus Clustering. Advances in K-means Clustering 2012; 155–175.
[11] Steinley D. Standardizing Variables in K-means Clustering. Classification, Clustering, and Data Mining Applications 2004; 53–60.
[12] Hosmer DW Jr, Lemeshow S. Applied Logistic Regression. John Wiley & Sons, 2004.
[13] Yao L, Zhong Y, Wu J, et al. Multivariable Logistic Regression And Back Propagation Artificial Neural Network To Predict Diabetic Retinopathy. Diabetes MetabSyndrObes 2019; 12: 1943–1951.
[14] Jahan NZ, Zahira Jahan N, Sasitharan null T. Customer Classification Of Discrete Customer Assets Data And Re-Ranking Of Classified Data. International journal of computer techniques; 7. Epub ahead of print 2020. DOI: 10.29126/23942231/ijct-v7i2p3.
[15] SPSS Software, https://www.ibm.com/analytics/spss-statistics-software (accessed 6 July 2021).
[16] Dingli A, Marmara V, Fournier NS. Enhancing Customer Retention Through Data Mining Techniques. Machine Learning and Applications: An International Journal 2017; 4: 01–10.
[17] Pan S, Qiao J, Zhu L. Application of Parallel Clustering Algorithm Based on R in Power Customer Classification. 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). Epub ahead of print 2019. DOI: 10.1109/icccbda.2019.8725760.
[18] Sokol O, Holý V. The role of shopping mission in retail customer segmentation. International Journal of Market Research 2020; 147078532092101.
[19] Clustering of electricity consumption behavior dynamics towards big data applications. Int J Recent Trends Eng Res 2018; 4: 102–106.