A Machine Learning Model for Customer Segmentation in a Telecom Company using K –Nearest Neighbor (KNN)

Israa Abdulrauof Othman
Assistant Professor
Software Engineering Program
Elnasr Technical College

ABSTRACT

Now a days many countries hosts a fluid and competitive telecommunication market and for a company to create .sustain customer value and increase economic efficiency, it needs to better understand its customers. The purpose of customer segmentation or clustering is to deliver actionable results for marketing, business planning and product development .This paper focus on customer segmentation using clustering algorithms on real data of a telecommunication company, the dataset is from IBM recourses. After choosing appropriate attributes for clustering, KNN clustering algorithm was used in order to create different customer segments. Moreover, the insights obtained from each segment were analyzed before suggesting marketing strategies for up- selling and better targeted campaigns.

General Terms

Machine Learning, Algorithms, K –Nearest Neighbor (KNN)

Keywords

Machine Learning, Nearest Neighbor (KNN), Customer Segmentation, Telecom Company

1. INTRODUCTION

Telecom companies gather different kinds of data to make sure their services answer customers' needs. So the service providers must keep the consumers happy and maintain their happiness to make sure that their business will stand out in this super-competitive market. Therefore, boosting customer experience to make marketing more effective is one of the key segmentation advantages in telecom.

1.1 Customer segmentation techniques

Customer segmentation groups similar customers to do an in-depth analysis of their behavior. Customers can be segmented by their industry, product tier or usage level.[1]employee size ,Revenue, geographical locations, andchannel their contact originated from, the date they became a customer, buyer persona and ideal customer profile. Grouping by these data points helps companies gain several segmentation advantages in telecoms, such as determining which types of customers won't like the specific products or services and identifying the ideal customer profiles. The purpose of such segmentation is to identify various groups within the company's target audience so that marketers can deliver valuable messaging and more targeted[3].

2. LITERATURE REVIEW

[1].Kim et. Al (2014) proposed a churn prediction model by examining the communication patterns among subscribers and considering a relationship between churners to non-churners. This paper highlighted that the relation among customer in the community can be the determinant of customer churn, not only the customer's characteristics,. This paper compared the predicting power with LR and MLP. [3][4]

[2]. Vafeiadis et. al. (2015) implemented performance comparison of several machine learning algorithms: (DT, ANN, SVM,LR, and NB). This study adopted cross-validation and boosting techniques, and various evaluation measures (accuracy,F-measure, recall, precision). This paper argued that boosted SVM had best accuracy (97%) and F-measure (84%), while the second-best model was ANN which had 94% accuracy and 77%. [5]

[3]. In Korea, many studies also studied related to customer churn, reflecting the intense competition in the Korean mobile communication market. Ahn et. al. (2006) studied conceptual model for churn in Korean mobile market. This paper investigated customer churn determinant (Customer status, Customer dissatisfaction, Switching cost, Service usage, and Customer-related variables), and tested the hypothesis with Logistic Regression. Call Billed amounts, drop rate, Number of complaints, Loyalty points, Customer status are determinedaccepted hypothesis [6]

[4]. F. Sabbeh et al(2018) ,present's a benchmark for the most widely used state of the arts for churn classification. The accuracy of the selected models was evaluated on a public dataset of customers in Telecom Company. Based on the findings of this study, ensemble – based learning techniques are recommended as both Random forest and Ad boost models gave the best accuracy[7].

[5]. Nurulhuda et al(2022) applied data mining techniques to the NPS dataset from a Malaysian telecommunications company in September 2019 and September 2020, analyzing 7776 records with 30 fields to determine which variables were significant for the churn prediction model. by developing a Linear Discriminant Analysis, K-Nearest Neighbors Classifier, Classification and Regression Trees (CART), Gaussian Naïve Bayes, and Support Vector Machine using 33 variables. As a results: Customer churn is elevated for customers with a low NPS. CART has the most accurate churn prediction (98%). However, the research is prohibited from accessing personal customer information under Malaysia's data protection policy. Results are expected for other businesses to measure potentialcustomer churn using NPS scores to gather customer feedback [8].

3. METHDOLOGY

3.1 K-Nearest Neighbors

The KNN machine learning method works on the distance between the data points and is based on the basic premise that similar data points are close to each other. It is widely used in image recognition, customer recommendation systems, and decision-making models. If neighbors are categorized in a certain way, the model will likely fall under the same category too based on the virtue of you being near to them. Conversely, those who are not near but farther away are likely to fall under a different classification than others. There can be exceptions (points belonging to a different category being near to points in a different category) and consequently, there can be errors in the predictions based on distance, especially around the decision boundaries. KNN model can be implemented and check how accurately it can predict the classification of points. Think of these data points as points in an n-dimensional space and the KNN algorithm predicting their categories based on the distance between the uncategorized data points and their k nearest neighbors, where k is a parameter.[9] The model applied to dataset of a telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups. If demographic data can be used to predict group membership, the company can customize offers for individual prospective customers. This is a classification problem. That is, given the dataset, with predefined labels, a model is built to be used to predict the class of a new or unknown case.

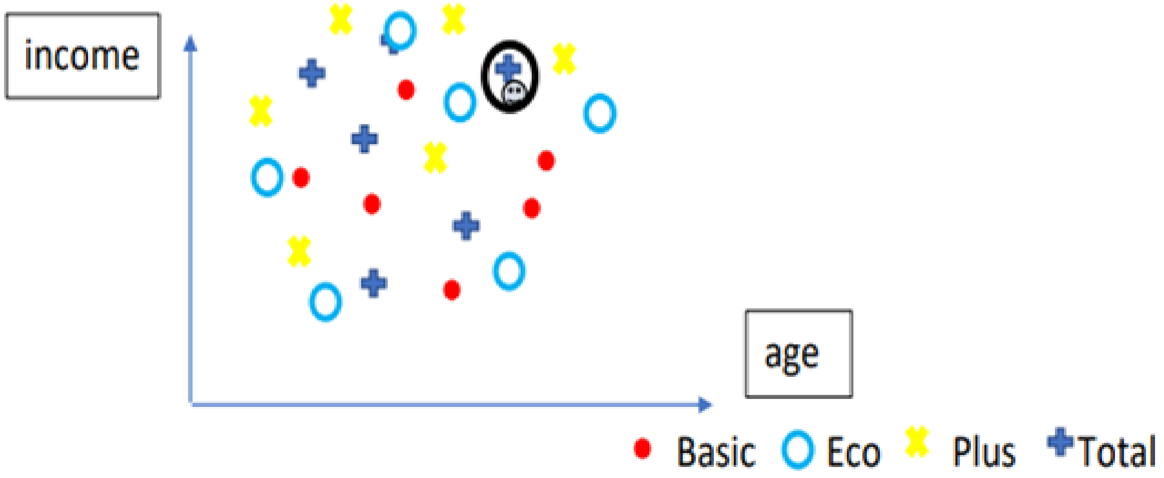


Fig 1: Customer Segments

3.2 Experimentation and Data analysis

The environment which used for this research is as follows:

3.2.1. jupyter notebook:

Jupiternotebook is an environment that totally runs on cloud. It allows you to write and execute any python code through the browser and is highly suitable for data science and machine learning.

3.2.2 Python:

Python is one of the most popular programming language at present. It supports multiple paradigms like object oriented, structure oriented, and function oriented programming. It is a general-purpose high-level interpreted and programming language.

3.2.2.1 Numpy:

Numpy is an open source python library that is used to work with arrays. It stands for Numerical Python. It has all the mathematical functions for working with domains like matrices and linear algebra . Numpy library can be imported using the statement "import numpy".

3.2.2.2 Pandas:

Pandas is an open source python library that is used to work with data sets. It has number of functions for exploring, analyzing, cleaning and manipulating data. The pandas library can be imported using the statement "import pandas".

3.2.2.3 Sklearn:

Sklearn (Scikit-learn) is one of the most useful python library that can be used for machine learning. Sklearn includes many functions for machine learning and modelling such as Clustering, Classification, Dimensionality Reduction, and Regression. Sklearn can be installed using the pip command "pip install -U scikit-learn".

3.2.2.4 Matplotlib:

Matplot is an open source graph plotting library used for visualization. It is generally written in python, some of the segments can be compatible with Java script, C, and Objective-C. We can import the library with the statement "import matplotlib".

3.3 prediction and Results

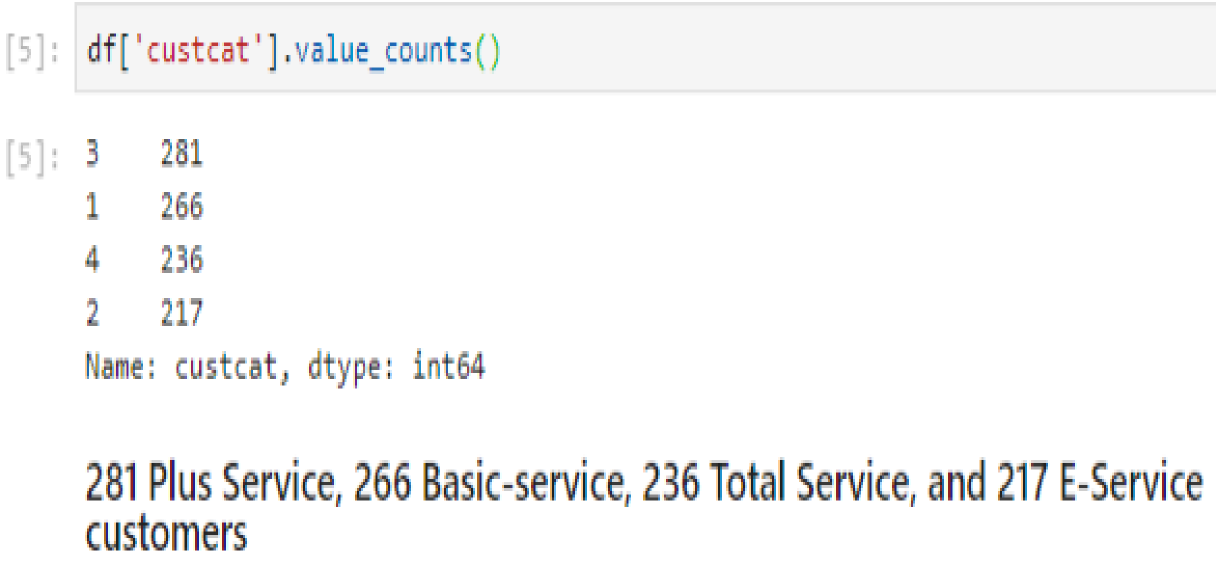
The data that used for this research is provided as open source by IBM.

The target field, called custcat, has four possible values that correspond to the fourcustomer groups, as follows:Total Service,Plus Service, E-Service, and Basic Service. This paper objective is to build a classifier, for example using the rows 0 to 7, to predictthe class of row 8.a specific type of classification called K-nearest neighbor is used.Just for sake of demonstration, let’s use only two fields as predictors - specifically,Age and Income, and then plot the customers based on their group membership.The k-nearest-neighbors algorithm is a classification algorithm that takes a bunch of labelled pointsand uses them to learn how to label other points.This algorithm classifies cases based on their similarity to other cases.In k-nearest neighbors, data points that are near each other are said to be “neighbors.”K-nearest neighbors is based on this paradigm: “Similar cases with the same class labelsare near each other.”Thus, the distance between two cases is a measure of their dissimilarity.There are different ways to calculate the similarity, or conversely, the distance ordissimilarity of two data points.In a classification problem, the k-nearest neighbor’s algorithm works as follows:

1. Pick a value for K.
2. Calculate the distance from the new case (holdout from each of the cases in the dataset).
3. Search for the K observations in the training data that are ‘nearest’ to the measurementsof the unknown data point.

And 4, predict the response of the unknown data point using the most popular response value fromthe K nearest neighbors.

, K in k-nearest neighbors, is the number of nearest neighbors to examine.Actually, since its nearest neighbor is Blue, that means captured the noise in thedata, or chose one of the points that was an anomaly in the data.A low value of K causes a highly complex model as well, which might result in over-fittingof the model.It means the prediction process is not generalized enough to be used for out-of-sample cases.Out-of-sample data is data that is outside of the dataset used to train the model.In other words, it cannot be trusted to be used for prediction of unknown samples.It’s important to remember that over-fitting is bad, as we want a general model that worksfor any data, not just the data used for training.On the opposite side of the spectrum, if we choose a very high value of K, suchas K=20, then the model becomes overly generalized.The general solution is to reserve a part of the data for testing the accuracy of themodel.Once it is done so, choose k =1, and then use the training part for modeling, and calculatethe accuracy of prediction using all samples in the test set.Repeat this process, increasing the k, and see which k is best for the model.In our case, k=4 will give us the best accuracy.Nearest neighbors analysis can also be used to compute values for a continuous target



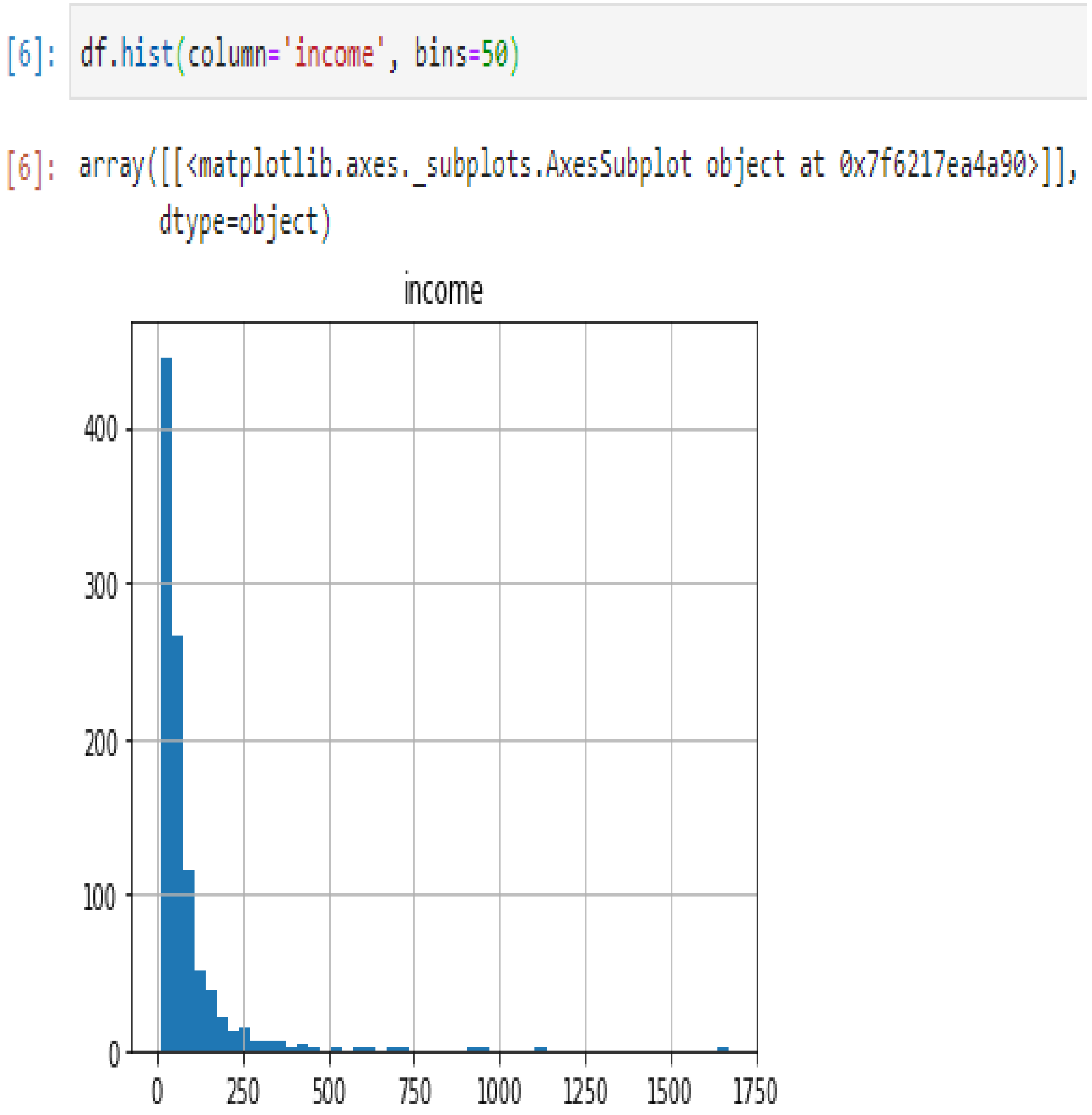
```
[5]: df['custcat'].value_counts()
```

```
[5]: 3    281
     1    266
     4    236
     2    217
     Name: custcat, dtype: int64
```

281 Plus Service, 266 Basic-service, 236 Total Service, and 217 E-Service customers

Fig 2:Class’s Value Count

Then calculate total number of customers in each classIn order to make the form of the data clearer, the graph of the customer data was drawnas follow:

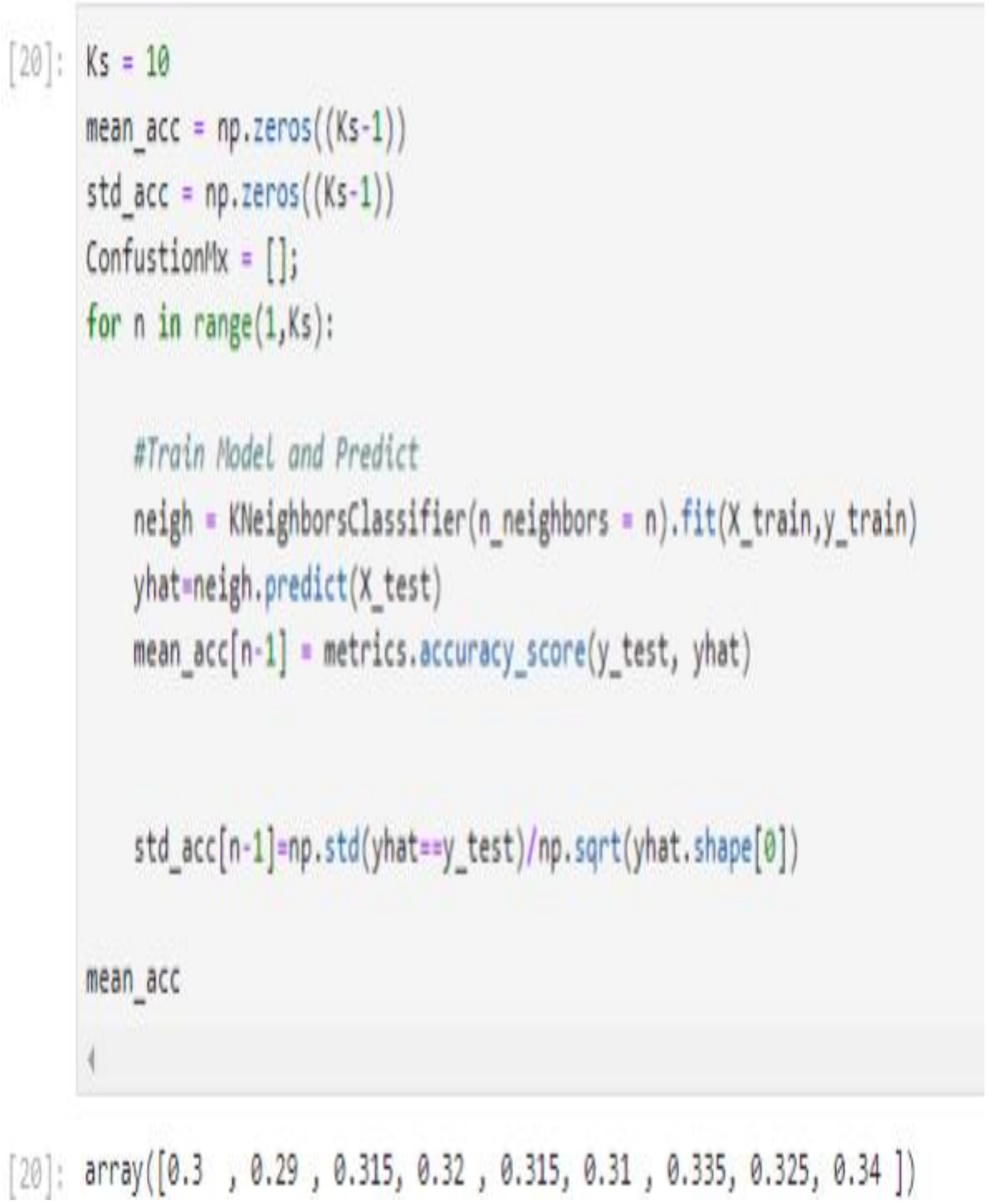


```
[6]: df.hist(column='income', bins=50)
```

```
[6]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f6217ea4a90>]],
           dtype=object)
```

Fig 3: data histogram

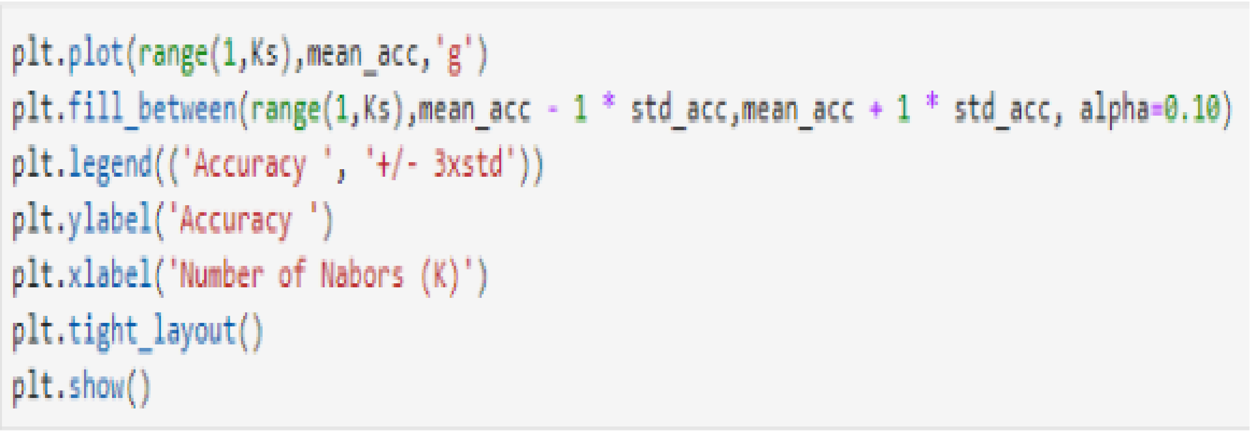
Then the variableK is entered with several values, including 10.



```
[20]: Ks = 10
      mean_acc = np.zeros((Ks-1))
      std_acc = np.zeros((Ks-1))
      ConfustionMx = [];
      for n in range(1,Ks):

          #Train Model and Predict
          neigh = KNeighborsClassifier(n_neighbors = n).fit(X_train,y_train)
          yhat=neigh.predict(X_test)
          mean_acc[n-1] = metrics.accuracy_score(y_test, yhat)


          std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])

      mean_acc

[20]: array([0.3  , 0.29 , 0.315, 0.32 , 0.315, 0.31 , 0.335, 0.325, 0.34 ])
```

Fig 4: Train and Test Data Splitting

As shown in fig 5 the accuracy of the model was plotted compared to K different values to determine the highest Accuracy at any valueof K,The highest accuracy was reached whenK=9.



```
: plt.plot(range(1,Ks),mean_acc,'g')
  plt.fill_between(range(1,Ks),mean_acc - 1 * std_acc,mean_acc + 1 * std_acc, alpha=0.10)
  plt.legend(('Accuracy ', '+/- 3xstd'))
  plt.ylabel('Accuracy ')
  plt.xlabel('Number of Nabors (K)')
  plt.tight_layout()
  plt.show()
```

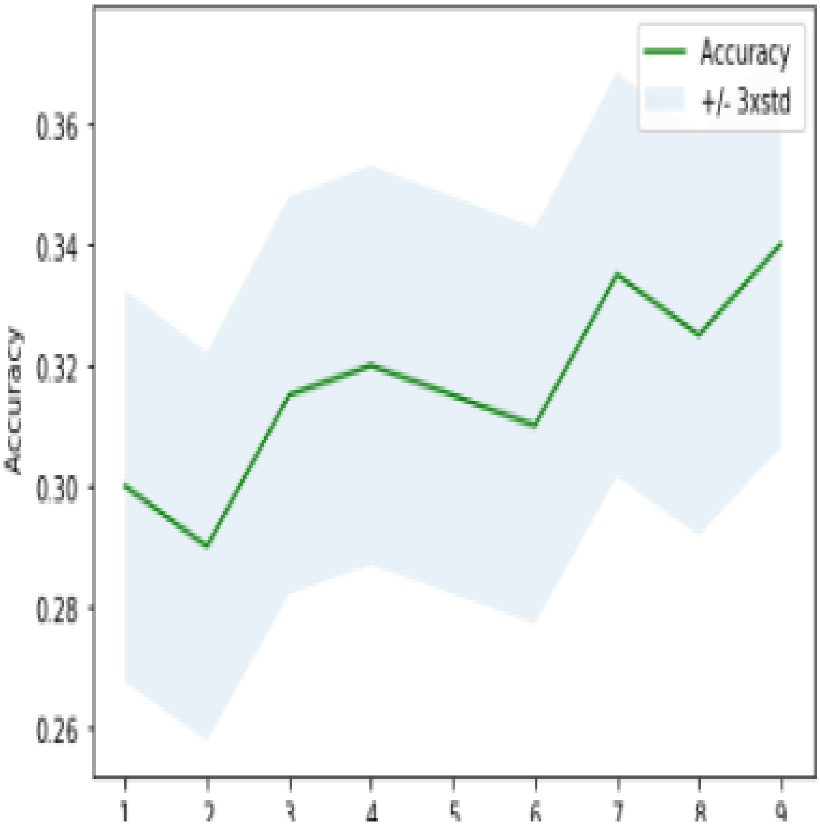


Fig 5: Model Accuracy

4. CONCLUSION AND FUTURE WORKS

This paper shown that through the use of customer segmentation, a telecommunication company could easily market its customers with right services and products. On the other hand, this also helps in offering tailored packages, bundles and offers for customers. In this way, it becomes easier for company officials to create marketing campaigns from scratch for specific customer segments instead of the whole customer base. A classifier model is built to classify customers to four possible values that correspond to the four customer groups, as follows: Total Service, Plus Service, E-Service, and Basic Service, that will help company to increase selling with successful customer targeting. Further research is required into how to improve the classification accuracy of marginal data which fall outside the regions of representatives.

5. ACKNOWLEDGMENTS

I would like to thank IBM for rendering help under IBM Academic Initiative.

6. REFERENCES

[1] Jain S., Sharma N.K., Gupta S., Doohan N. (2018) Business .Strategy Prediction System for Market Basket Analysis. In: Kapur P., Kumar U., Verma A. (eds) Quality, IT and Business Operations. Springer Proceedings in Business and Economics. Springer, Singapore. 2017. DOI https://doi.org/10.1007/978-981-10-5577-5_8

[2] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K., “Customer churn prediction in the telecommunication sector using a rough set approach,”Neurocomputing, 237, pp. 242-254, 2017.

[3] TASIL, https://tasil.com/insights/segmentation-advantages-in-telecom/23/8/2022

[4] Kim, K., Jun, C. H., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. Expert Systems with Applications, 41(15), 6575-6584.

[5] H. Wang.: Nearest Neighbours without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N.Ireland, UK (2002)

[6] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas,K.C.(2015).A comparisonof machine learning techniques for customer churn prediction. Simulation Modelling Practiceand Theory, 55, 1-9.

[7] Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10-11), 552-568.

[8] Akhil Sharma,Published inPublished in Data Science https://medium.com/data-science-on-customer-churn data/k-nearest-neighbors-knn-on-customer-churn-data-40e9b2bb9266

[9] Customer churn prediction
for telecommunication industry: A Malaysian Case Study [version 1; peer review: 2 approved] Nurulhuda Mustafa1, Lew Sook Ling 2, Siti Fatimah Abdul Razak 2 F1000Research 2021, 10:1274 Last updated: 13 JUN 2022