

PREDICTION OF ATTRITING CUSTOMERS

CA1 – PATTERN RECOGNITION SYSTEMS

*GROUP MEMBERS* *STUDENT ID*

Aniket Mohan Arasanipalai A0198481N

Akshay Sachdeva A0198532W

Contents

[**1.** **BUSINESS PROBLEM BACKGROUND** 2](#_Toc20436903)

[**2. OBJECTIVES AND SUCCESS MEASUREMENTS** 3](#_Toc20436904)

[**2.1 OBJECTIVES** 3](#_Toc20436905)

[**2.2 SUCCESS MEASUREMENTS** 3](#_Toc20436906)

[**3. SOLUTION** 4](#_Toc20436907)

[**3.1 TOOLS USED** 4](#_Toc20436908)

[**3.2 TECHNIQUES** 4](#_Toc20436909)

[**3.3 MODEL SUMMARY** 8](#_Toc20436910)

[**4. CONCLUSION** 9](#_Toc20436911)

[**5. REFERENCES** 9](#_Toc20436912)

# **BUSINESS PROBLEM BACKGROUND**

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company. Organizations nowadays are, always interested in identifying segments of these customers because the price for acquiring a new customer is usually higher than retaining the old one. For example, if Netflix knew a segment of customers who were at risk of churning, they could proactively engage them with special offers instead of simply losing them.

In this post, we will create a simple customer churn prediction model using a bank dataset. We chose 3 standardized models and 1 ensemble model to model churned customers, pandas for data crunching and matplotlib for visualizations. We will do all of that above in Python.

The code can be used with another dataset with a few minor adjustments to train the baseline model. We also provide few references and give ideas for new features and improvements.

# **2. OBJECTIVES AND SUCCESS MEASUREMENTS**

## **2.1 OBJECTIVES**

The objective of this project is to create a customer churn model which can provide high accuracy and least error to predict customers’ attrition

With the vast majority of organizations these days, looking to minimize losses, and with so many factors and categories to keep track of, it becomes easy if they are able to identify the customer which may or may not require attention, in order to minimize the losses.

## **2.2 SUCCESS MEASUREMENTS**

It is vital to understand the success of the model based on a few measures.

There are a few key measures of our model:

1. Whether the model was able to provide a good accuracy
2. Whether the model were not error prone
3. Whether the model could be tweaked easily for improvement

# **3. SOLUTION**

## **3.1 TOOLS USED**

For this project, we have used Python 3 as our code benchmark.

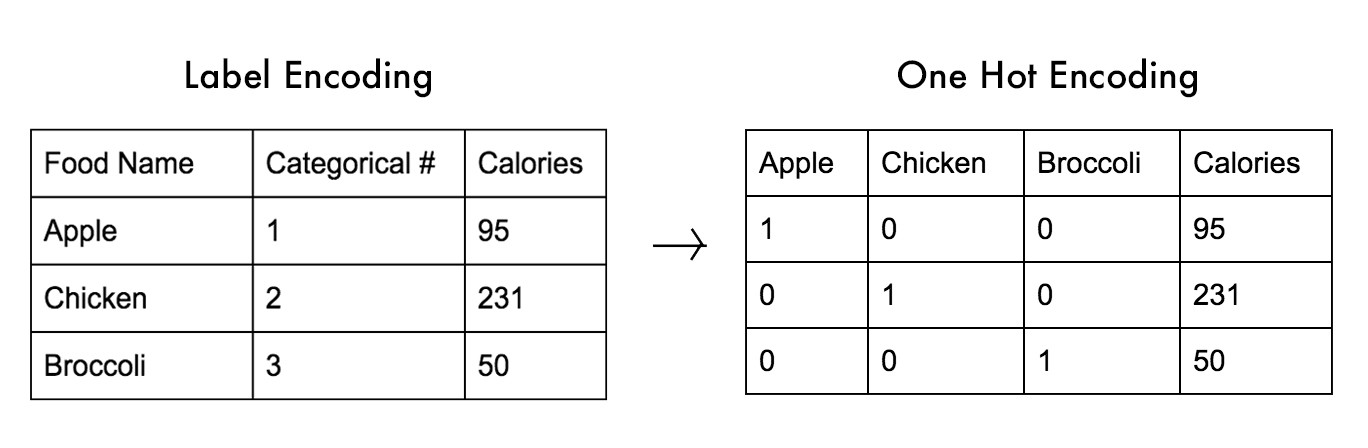
## **3.2 TECHNIQUES**

**Data Pre-Processing:**

The shape of the dataset was initially 7043 \* 31. To obtain the data in a clean format, we pre-processed the data through various steps.

Firstly, we excluded the columns like ‘CustomerID’, ‘Partner’, and ‘Contact’ among others (in all 5) bringing down the data to 26 columns. Furthermore, we split the categorical variables and obtain only the first element of the values, i.e. ‘Yes’ becomes ‘Y’ and ‘No’ becomes ‘N.

Post this, we convert the aforementioned categorical values to numerical data and perform *one hot encoding* on the data. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. To further understand, let’s take the help of an example –



The categories were formerly rows, but now they’re columns. In other words, an additional binary column for each category is created.

Once we have the pre-processed the data, we split the data set into training (70%) and test(30%) data and accordingly run our models on them.

**Modelling Techniques:**

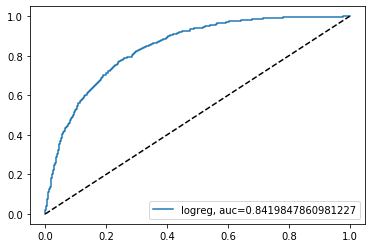
1. **Logistic Regression**

Logistic regression, a classification algorithm, is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

In our case, we have made use of the algorithm, to predict churn (0 being no, 1 being yes)

***Performance of the Model:***

For this model we have obtained, an accuracy of ~=0.80 on the testing data, with a mean square error of 0.20.



1. **KNN(K Nearest Neighbour)**

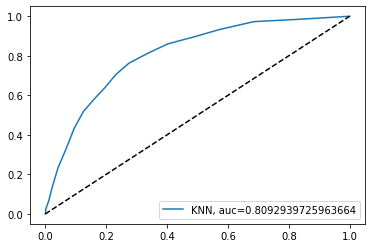
The k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.

***Performance of the Model:***

For the KNN mode, we have obtained, an accuracy of 0.78 with a mean square error of 0.20.

***Insight:***

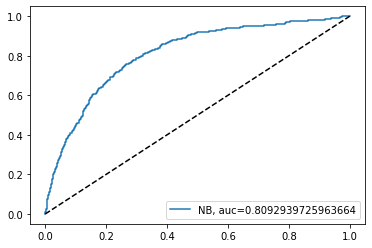
Based on the above Logistic Regression and KNN results, the results do not vary a lot. Historically, the improvement between these algorithms has not been seen too significant in terms of accuracy or reduction of mean square error.



1. **Naïve Bayes**

Naive Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting.

For the Naïve Bayes algorithm, we have obtained an accuracy of 0.76, while a mean square error of 0.23 is obtained.



1. **Random Forest (Ensemble Method)**

Random forests or random decision forests are an *ensemble learning method* for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Due to its general use in classification models, and the success which its being used with, we have utilized the same for our churn prediction.

On using Random Forest, we have seen an improvement in our model with accuracy ~=0.80 and mean square error remaining the same at 0.20

## **3.3 MODEL SUMMARY**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy (%) | Mean Square Error |
| Logistic Regression | 79.98 | 0.20 |
| KNN | 78.02 | 0.20 |
| Naïve Bayes | 76.06 | 0.23 |
| Random Forest | 79.55 | 0.20 |

# **4. CONCLUSION**

Based on the machine learning models run for the churn prediction, we see that Random Forest and Logistic Regression provide fairly better accuracy and lower MSE than the other two algorithms – KNN and Naïve Bayes.

Therefore, we could go further with improving the existing algorithm by tweaking the parameters to obtain an even higher accuracy.

# **5. REFERENCES**

1. [A Survey on Customer Churn Prediction using Machine Learning Techniques](https://www.researchgate.net/publication/310757545_A_Survey_on_Customer_Churn_Prediction_using_Machine_Learning_Techniques)] — This paper reviews the most popular machine learning algorithms used by researchers for churn predicting.
2. [WTTE-RNN-Hackless-churn-modeling](https://ragulpr.github.io/2016/12/22/WTTE-RNN-Hackless-churn-modeling) — Event based churn prediction.