Part A (Predicting Bank Marketing Campaign Outcomes)

This part of the assignment is concerned with predicting the outcome of direct bank marketing campaigns (phone calls) of Portuguese banking. The dataset (Bank.zip) contains 17 attributes for which outcomes of subscribing to a term deposit (yes/no) are known. You are required to build models using the K-Nearest Neighbors (KNN) and Naïve Bayes (NB) algorithms.

a)**Explain the KNN and Naïve Bayes Algorithms [10 marks]**

In your own words, explain how each of the KNN and Naïve Bayes algorithms work.

## KNN Algorithm

K Nearest Neighbours is also a form of Supervised Machine learning and it classifies a data point based on how its neighbours are classified. Simply put, classification is based on a feature similarity measure by looking at the characteristics of the nearest neighbours. The distance between the data points is calculated and the data points are weighted according to distance. The data points that are closest to the point in question have the highest weight in determining the class. There are at least 3 methods that can be used to calculate distance. These are Euclidean distance, Manhattan distance and q norm distance. Where data is nominal, Euclidean distance should not be used.

In the Textbook; Introduction to Data Minin, Tan et al, 2019, explain the KNN Algorithm as an equation calculating distance for datapoint z which is the datapoint in question. Data points that are located far away from z have a weaker impact on the classification of z in comparison to data points that are closer to z, Tan et al, 2019.

A white paper with black text

Description automatically generated

Figure 1: Theoretical KNN Algorithm sourced by Tan et al, 2019.

The process of selecting K is very critical in the success of accurate classification of the KNN model. If we choose a value of K that is too small, z will become sensitive to noise. If we choose a value of K that is too large, the neighbourhood will include points from other classes that are not z.

Using a practical example below, Figure 1.1, we are trying to determine if the orange apple below belongs to the black apples of the white apples. K Nearest Neighbour assumes a value for K. Essentially K is how many neighbouring datapoints shall we examine to classify a datapoint. In this example let’s assume K = 5. We will look at the 5 nearest apples to the orange apple. Since 3 out of 5 apples with the green circle of neighbouring apples are white, we will presume that the orange apple is in fact meant to be a white apple.

A screenshot of a computer screen

Description automatically generated

Figure 1.1: Probability of No Purchase vs Purchase with Discount and Free Delivery on a Holiday

## Naïve Bayes Algorithm

Naïve Bayes is a Machine Learning method that is a part of Supervised Learning and is used to Classify objects. Examples of its use can be shown in weather predictions and medical diagnosis. Naïve Bayes is suitable for Continuous Data. The advantages of this model are it is simple and fast to implement and it is robust enough to deal with unknown/missing values

Naïve Bayes works on the principles of conditional probability which was derived from Bayes Theorem. As show in the formula below, the Bayes Theorem gives the conditional probability of an event A given another event B has occurred.

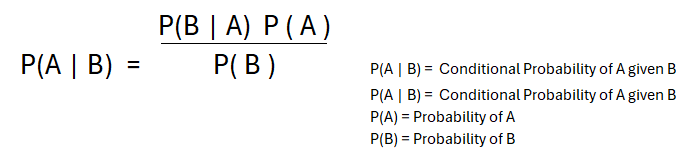


Figure 1: Bayes Theorem Conditional Probability Formula

The Naïve Bayes Algorithm uses Frequency and Likelihood tables to determine the conditional probability (Bayes Theorem) of an event A given event B.

To better explain how the Naïve Bayes Algorithm works we can use the shopping example dataset below:

A white rectangular table with black text

Description automatically generated

Table 1: Shopping Dataset

Below is a calculation of the Frequency Table and Likelihood Table.

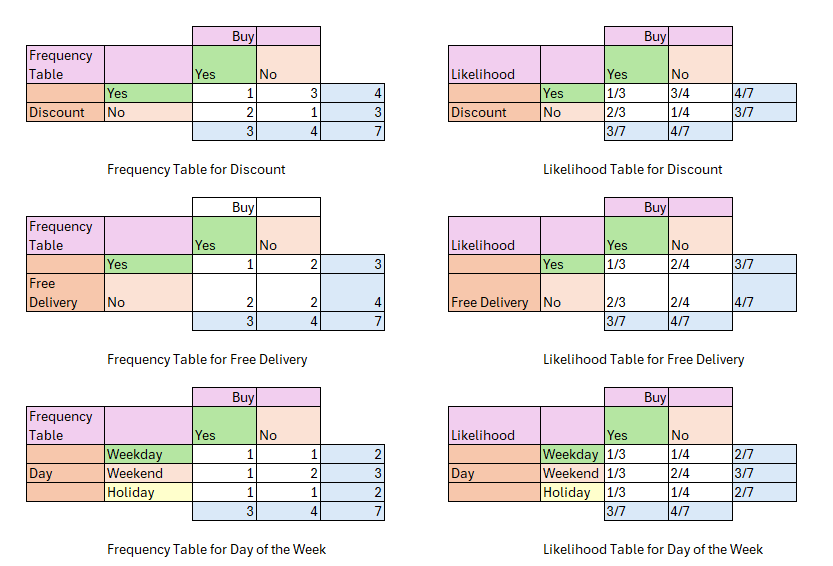


Figure 1.2: Frequency Table and Likelihood Table for Shopping Dataset

From the Frequency and Likelihood Tables in Figure 1.2, we can determine the probability of whether there will be No Buys on a Holiday with a Discount and Free Delivery vs Buys on a Holiday with a Discount and Free Delivery. These tables tell us how many times specific combinations of certain Buy and No Buy events have occurred and under what conditions ie: Weekend, Weekday, Holiday, Discount, No Discount, Free Delivery and No Free Delivery.

As an example, we can then determine the probability of No Buy with Discount and Free delivery on a Holiday and the probability of Buy with Discount and Free delivery on a Holiday. This is shown in Figure 1.4 below.

A screenshot of a computer

Description automatically generated

Figure 1.4: Probability of No Purchase vs Purchase with Discount and Free Delivery on a Holiday

As per Figure 1.4, the probability of No Buy with Discount and Free delivery on a Holiday is 76.60% while the probability of Buy with Discount and Free delivery on a Holiday 22.70%.

To finalise the Naïve Bayes Algorithm, the probabilities then need to be Normalised to determine the likelihood of the events occurring. This is how the Naïve Bayes Model classifies the data. The Figure below shows the normalisation process.

A screenshot of a calculator

Description automatically generated

Figure 1.5: Normalising the Probabilities of the Shopping Dataset

Figure 1.5 tells us the likelihood of Purchase with Discount, Free Delivery on a holiday is 22.86% and the likelihood of No Purchase with Discount, Free Delivery on a Holiday is 77.14%. From this, we can conclude that the shopping data tells us that on average, customers will not buy on a holiday with a discount and free delivery.

b) **Perform Exploratory Data Analysis (EDA) [10 marks]**

Perform EDA and describe your dataset. Explain any pre-processing and data manipulation tasks you performed to prepare your dataset for building your models. Note: No grade will be given for presenting plots/tables without explanation.

The data set contains information related to a direct marketing campaign carried out by a Portuguese bank during May 2008 and November 2010. The data describes the age, job, marital status and education about the customers. It also specifies whether the customers are in credit default, average yearly balance of the account, their housing and personal loan details, and various details regarding how the bank contacted the customer for marketing purposes.

Figure 1.6 below shows the count of observations, features and data types for the bank dataset. It has 4521 datapoints and 17 features. The datapoints are a combination of numeric and categorical information with the following being numeric:

* age,
* balance,
* last contact day,
* last contact duration,
* number of contacts during campaign,
* previous contacts

While the remaining features are categorical

* job,
* marital status,
* education,
* credit in default,
* housing loan,
* personal loan,
* contact type,
* contact month,
* past days,
* previous outcome
* sub term deposit

A screenshot of a computer

Description automatically generated

Figure 1.6: Banks data set instances, observations and data types

Figure 1.7 shows that the dataset has no rows with missing values no does it contain duplicate rows.

A screenshot of a computer

Description automatically generated

Figure 1.7: Duplicate rows and Missing Values

After examining the numerical data to check for outliers, the following columns have been identified as having outliers with a z-score of 1.5 and a z-score of 3. Figure 1.8 shows the age, balance, last contact day, last contact duration, num of contacts during the campaign and previous contacts have datapoints that fall within the z-score threshold of 1.5. This threshold is not typically considered extreme. On the other hand, there has been an indication that there are some datapoints in the age, balance, last contact duration, num of contacts during the campaign and previous contacts columns that have a z-score threshold of 3 meaning that there are some datapoints with extreme values in these columns and special considerations might be necessary. Specifically, 391 rows have been identified as being with this z-score threshold of 3. As a result, these rows will be removed from the dataset.

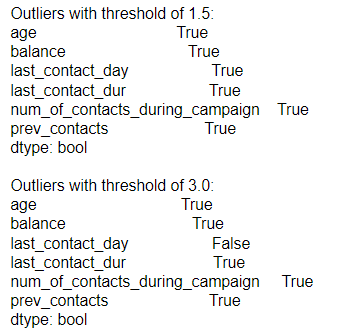




Figure 1.8: Outliers in the Bank Dataset with a z-score of 1.5 and 3.00

c) **Feature Selection and Analysis [10 marks]**

Identify the most influential features in classifying this dataset using an appropriate method. Explain the process of the chosen feature selection method and use the **top five features** for building your models. Use a breakdown analysis for selected features by class and describe their distribution using appropriate plots.

d) **Independence Assumption in Naïve Bayes [5 marks]**

Discuss the independence assumption between the features in the Naïve Bayes algorithm and support your answer with respect to the selected features.

e) **Naïve Bayes Model Building and Evaluation [10 marks]**

Run the Naïve Bayes algorithm with the GaussianNB implementation for the selected features. Provide evaluation metrics, including the confusion matrix, showing the performance of the NB model. Discuss the results.

f) **KNN Model Building and Evaluation [10 marks]**

Fit a KNN model for a range of K values. Provide and examine the confusion matrix. Generate and provide a classification report showing precision, recall, F1 score, and overall accuracy to evaluate your model performance. Discuss the results.

g) **Model Comparison [5 marks]**

Compare the performance of your KNN and NB models. Discuss your findings.

Part B: Exploring Artificial Neural Networks

In this part, you are required to explore various architectures for building an Artificial Neural Network (ANN). Use the 10-fold cross-validation option for testing.

**Tasks**

**a) Activation Function and Learning Rate in MLP [5 marks]**

Explain the role of an activation function and learning rate in building a Multilayer Perceptron (MLP).

**b) Baseline Model with MLPClassifier [5 marks]**

Use the sklearn.MLPClassifier with default parameter values and a single hidden layer with 𝑘𝑘 neurons (𝑘𝑘 ≤ 25). Determine and report the best number of iterations that gives the highest accuracy. Use this classification accuracy as a baseline for comparison in later parts of this question.

**c) Tracking Loss Value [5 marks]**

Enable the loss value to be shown on the training segment and track the loss as a function of the iteration count. Explain any observed discrepancies between loss value and error value over consecutive iterations.

d) **Experimenting with Two Hidden Layers [10 marks]**

Experiment with two hidden layers and experimentally determine the split of the number of neurons across each of the two layers that gives the highest classification accuracy. In part 1, you had all k neurons in a single layer, in this part you will transfer neurons from the first hidden layer to the second iteratively in step size of 1. Thus, for example in the first iteration, the first hidden layer will have k-1 neurons whilst the second layer will have 1, in the second iteration k-2 neurons will be in the first layer with 2 in the second and so on. Summarise your classification accuracy results in a 25 by 2 table with the first column specifying the combination of neurons used (e.g., 12, 13) and the second column specifying the classification accuracy.

**e) Explaining Accuracy Variation [5 marks]**

From the table created in part d, you will observe the accuracy variation with the split of neurons across the two layers. Give explanations for some possible reasons for this variation.

**f) Comparing MLP Classifier Performance [5 marks]**

Compare the performance of the MLP Classifier with other classifiers on your dataset in part A. Choose the best-performing model and explain why you chose it. Discuss your findings from the experiments and provide your opinion on these classifiers.

**Report Resentation [5 marks]**

**Submission Instructions**

Ony one submission per group is required. Please submit the following two files **separately** as part of your assignment:

1. Python Notebook (.ipynb)

2. Report File (PDF Format)

**Note**: the report should focus on presenting your findings and insights. Please refrain from including the code file in your report, as including code in the report will result in a penalty.

TO DO:

* EDA – changed ‘pdays’ or ‘past\_days’ to categorical from numeric
* Check z-score for balance, last\_contact\_dur to better represent overall data set