Name: \_\_\_\_Sri Vamshi Polela\_\_\_\_\_\_\_\_

CSDA 5430 Predictive Analytics

Week 4 Assignment

1. Probability and odds. In the personal loan example on chapter 10 PPT slides 13 ~15, the following predictions on test data (observations from 71 to 80) can be received.

A screenshot of a computer

Description automatically generated

* **What is the predicted probability that customer 189 (highlighted) would be a loan acceptor?**
* The predicted probability is 0.5540 from the given table.
* **What are the odds of the customer being a loan acceptor?**
* odds of the customer being a loan acceptor = = = 1.2422
* **What are the odds of the customer being a non-acceptor?**
* odds of the customer being a non-acceptor = = = 0.8051
* **When default probability** **threshold value 0.5 is used for classification, what is the equivalent threshold value in odds?**
* We know that odds =
* Probability threshold value is 0.5
* Odds = = 1
* The threshold value in odds is 1.
* **Will the same result be received if classification is performed with equivalent threshold odds value for the customer? Explain.**
* **The relationship between threshold probability (​p) and threshold odds value is:**

odds =

* **If the threshold odds value is calculated from a specific threshold probability and is set to be equivalent, the classification results will be the same.**

1. Extend the personal loan example on chapter 10 PPT slides 13 ~ 15 and fit a logistic regression model with two predictors: *Income* & *Age*.

* Write the estimated equation that associates the loan acceptance of a customer with the two predictors in three formats:
* **The logit as a function of the predictors**
* **The odds as a function of the predictors**
* **The probability as a function of the predictors**
* Consider a customer whose annual income is $141,000 and age is 42. From your logistic regression model, estimate the following quantities and classification for this customer:
* The logit of being a loan acceptor

= -0.89167

* The odds of being a loan acceptor

=

= = 0.4099705

* The probability of being a loan acceptor

=

=

= = 0.2907

* The classification of the customer (use probability threshold = 0.5)
* From the above probability (0.2907) which is below probability threshold classification of the customer is non-acceptor.

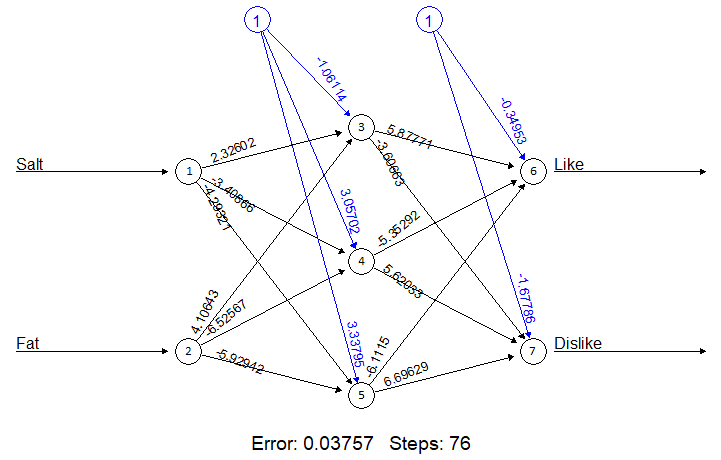
1. **Explain to your fellow coworkers the concept of neural network algorithm in predictive analytics. Discuss its black box nature and compare it with logistic regression algorithm. Use the example of personal loan acceptance in chapter 10 PPT to support your discussion.**

* **Neural networks**, inspired by the human brain, are complex machine learning models comprising interconnected layers of nodes (neurons) that process input data to yield output.
* With an architecture involving input, hidden, and output layers, each connection between nodes is assigned a weight learned during training.
* Activation functions introduce non-linearity, enabling the capture of intricate data relationships.
* The term **"black box"** highlights the lack of transparency in understanding how the network arrives at its decisions or predictions due to its complex and often non-linear structure. The internal processes of a neural network, especially in deep architectures with many layers, can be challenging to interpret, making it seem like a closed system or a "black box" in terms of its decision-making mechanisms.
* **In comparison**, logistic regression is simpler and easier to understand. It shows a direct connection between input features and predictions, making it more transparent. The trade-off is that neural networks can handle more complex patterns, while logistic regression keeps things straightforward and easier to interpret.

**Example:**

* Logistic regression helps us understand how things like income and age affect the chance of accepting a loan – it's like looking at each factor separately.
* A neural network can see complex connections between these factors, but it's like a mystery box – we might not fully grasp how it's making decisions because of its complexity.
* Choosing between logistic regression and a neural network depends on what we need – if we want a clear explanation, we might go with logistic regression, but if we need to catch more complex patterns, the neural network could be better. It's about finding the right balance.

1. Using the tiny neural network model presented on chapter 10 PPT slide 19, shown below, write the activation function for each hidden node (3, 4, & 5). For a new record: *Fat = 0.1* and *Salt = 0.5*, show the calculation for node 6 and 7 (from input layer, to hidden layer, to output layer). Use 4 decimal digits in calculation and results. What is the classification of this new record?



**Activation Function:**

**A math equation with a number of symbols

Description automatically generated with medium confidence**

**For Node 3:**

=

=

=

=

= 0.6254

* **For Node = 4**

=

=

=

= 0.6682

* **For Node = 5**

=

=

=

=

= 0.6452

* **Output Layer**
* **For Node 6**

=

=

=

= 0.01486

* **For Node 7**

=

=

=

=

= 0.9844

1. The dataset *ToyotaCorolla.csv* contains data on used cars (Toyota Corolla) on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on many attributes, including Price, Age, Kilometers, HP, and other specifications. The dealership would like to predict the price of a used Toyota Corolla car based on its specifications so the dealer will be able to predict the profit that the dealership will get for the used cars, assuming higher priced cars generate more profit.

In R, perform the following steps as discussed in Ch10 on the first 1,000 cars and the following variables: *Price, Age\_08\_04, KM, Fuel\_Type, HP, Met\_Color, Automatic, Doors, Quarterly\_Tax,* and *Weight*. A description of each variable is given at the end.

* Step 1: Collecting data (Discuss the business problem and how the data can support the analytics.)

**Business Problem:**

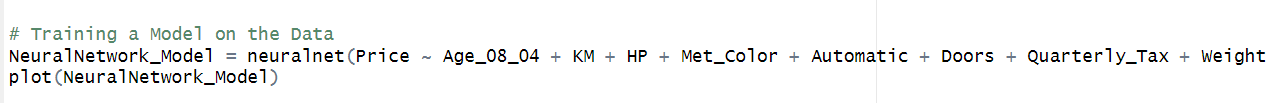
* Dealership wants to predict the price of used cars to predict the profit that the dealership will get for the used cars based on the predictors.

**Data Understanding:**

* The dataset "ToyotaCorolla.csv" comprises information on 1436 used Toyota Corolla cars for sale in the Netherlands in late summer 2004.
* The business objective is to predict the selling price of these cars based on key specifications, aiding the dealership in estimating potential profits.
* The selected variables, including Price, Age\_08\_04, KM, Fuel\_Type, HP, Met\_Color, Automatic, Doors, Quarterly\_Tax, and Weight, have been chosen to capture relevant features influencing car pricing.
* Step 2: Exploring data and preparing data (Use only the first 1000 rows of data and specified variables. Create dummies for categorical variable and normalize data. . Set a seed so the results can be reproduced.)
* After creating dummies for “fuel\_type” and using first 1000 rows of data

A screenshot of a computer

Description automatically generated

* Step 3: Training a model on the data (Use the function *neuralnet::neuralnet()* for the neural network algorithm. Remember to fit the model on the training data. Use default number for hidden neurons (hidden = 1) and plot the model.)
* 

A diagram of a graph

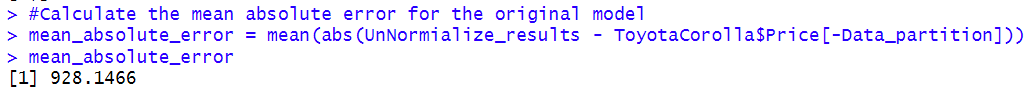
Description automatically generated

* **Step 4: Evaluating model performance (Make predictions on the testing data with the model. Make comparison between predicted results and observed value to evaluate the model performance. Do they correlate with each other? What is the mean absolute error between the two?)**
* I can see the correlation between the predicted results and observed value is 94% which is very high. They are highly correlated.
* I just wanted to see whether the correlation between predicted results and normalized test data And between predicted results and unnormalized test data. But both were the same.

A computer screen shot of a computer code

Description automatically generated

* The MAE for model with default hidden nodes (i.e., 1) is 928.1466.



* **Step 5: Improving model performance (Use a more complicated model with 4 hidden nodes to fit the same training data. Make prediction on the testing data with the new model. Compare performance between the new and the original model. Is there any performance improvement based on your comparison?)**
* After improving the model with hidden nodes = 4 and MAE is reduced to 896.9052

A screenshot of a computer program

Description automatically generated

* The improved neural network model, characterized by the addition of four hidden nodes, has enhanced performance compared to the original model.
* The evaluation results indicate a slightly lower Mean Absolute Error (MAE) of 896.9052 for the improved model, signifying improved accuracy in predicting Toyota Corolla prices on the testing data.
* Additionally, the correlation coefficient of 0.9456386 suggests a strong positive relationship between the predicted and actual prices.
* The neural network is built as shown below:

**A diagram of a network

Description automatically generated**

**Description of Variables in Toyota Corolla Dataset**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Price | Offer price in euros |
| Age\_08\_04 | Age in months as of August 2004 |
| KM | Accumulated kilometers on odometer |
| Fuel\_Type | Fuel type (*Petrol*, *Diesel*, *CNG*) |
| HP | Horsepower |
| Met\_Color | Metallic color? (Yes = 1, No = 0) |
| Automatic | Automatic (Yes = 1, No = 0) |
| Doors | Number of doors |
| Quarterly\_Tax | Quarterly road tax in euros |
| Weight | Weight in kilograms |