|  |  |
| --- | --- |
| Project milestone 6  Model Evaluation | Abstract  In this document, we compare and evaluate the performances and accuracies of the data models that we have created in the previous milestones. The Decision tree model and the Neural network model will be where our primary focus models, as they have been for the entirety of the project.  Dithapelo Huma and Maishibe thobela  BIN381 |

Table of Contents

[Decision Tree modelling technique 2](#_Toc71233687)

[Assumptions 2](#_Toc71233688)

[Challenges 2](#_Toc71233689)

[Model Environment 2](#_Toc71233690)

[Model Results 2](#_Toc71233691)

[Interpretation of 1st Confusion matrix 3](#_Toc71233692)

[Second decision tree 4](#_Toc71233693)

[Model results 4](#_Toc71233694)

[Interpretation of 2nd Confusion matrix 5](#_Toc71233695)

[Model conclusion 5](#_Toc71233696)

[Neural Network 7](#_Toc71233697)

[Model Description 7](#_Toc71233698)

[Challenges we had with our model 7](#_Toc71233699)

[Analysing the results of our original model 7](#_Toc71233700)

[Proof of evaluation 8](#_Toc71233701)

[Assessing the business success criteria 9](#_Toc71233702)

[R Code 9](#_Toc71233703)

# Decision Tree modelling technique

## Assumptions

* We start of by assuming that the data that is being used in the model is as clean as can be.
* There are no missing values in the attributes that we have chosen.
* And lastly, we can assume that all duplicate values have been removed.

## Challenges

**Categorical Data:** Data that was not numeric required us to either drop it or convert it into a factor data type. Such data made it difficult process and often gave us errors. Attributes such as the department name and occupation were not suitable for processing as they had a lot of uniwue values, which made it difficult to model as compared to numeric data

## Model Environment

**Attributes**

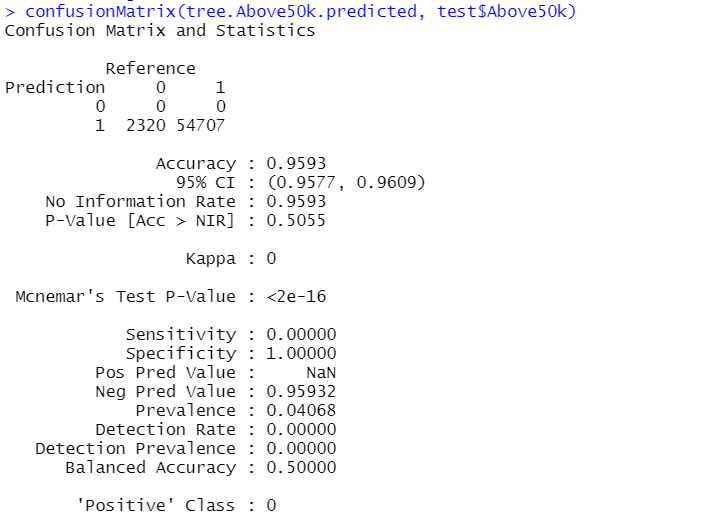
The following attributes were chosen for the decision tree attributes:

* Department Name
* Above50k
* SalaryType
* AgeGroup
* Years\_Worked

**Training and testing data**

* The dataset consists of 190 095 records.
* We used 70% of the data to train the model and the remaining 30% was used for testing purposes.

## Model Results



## Interpretation of 1st Confusion matrix

* The model was accurate in predicting that 54707 employees earned above R50k and those who worked at certain departments. It was however inaccurate in predicting the salary and department of 2303 employees
* The accuracy of the confusion matrix stands at 95%, and at a confidence interval of 95%, its accuracy lies between 0.9577 and 0.9609
* The Kappa value represents the agreement between two observers that are assessing the quality of their observations. With a Kappa value of 0, the internal comparison of the system did not agree with each other. If the Kappa value greater than 0.75, then we would conclude that the agreement would be excellent, and if it was below 0.4, it would show a poor agreement between observers
* With a P-value of 0.505, the is no statistical significance between the department that each employee works at and their salaries

**R code:**

#Libraries used

library(rpart)

library(rpart.plot)

library(dplyr)

library(data.tree)

library(caTools)

library(caret)

#Selecting Valuble Columns

colnames(rdataset)

rdata<-select(rdataset,Department.Name,Above50k, SalaryType, Agegroup, Years\_Worked )

set.seed(999)

sample = sample.split(rdata$Above50k,SplitRatio = 0.70)

train = subset(rdata, sample==TRUE)

test = subset(rdata,sample==FALSE)

#Training decision tree as a classifier

tree<- rpart(Above50k~., data = train)

#predictions

tree.Above50k.predicted<- predict(tree,test,type = 'class')

#evaluating the module with the confusion matrix

confusionMatrix(tree.Above50k.predicted, test$Above50k)

#decision tree

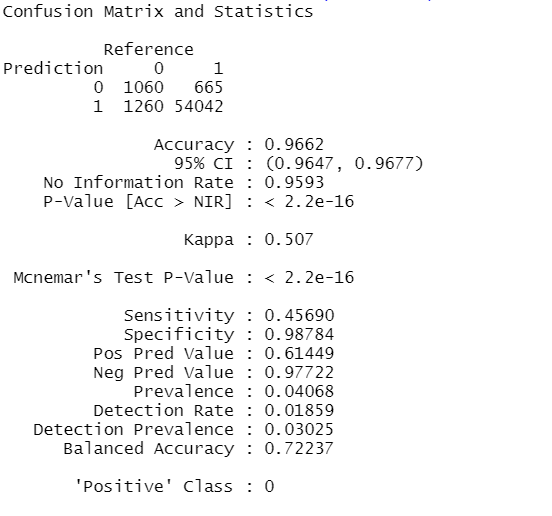
prp(tree)

## Second decision tree

A second decision tree was created to test the accuracy with different attributes. The attributes that were chosen for this tree are:

* Above50k
* DepRarity
* Agegroup
* SalaryType
* Years\_Worked

## Model results



## Interpretation of 2nd Confusion matrix

* This model compared employees who earned above R50k with the amount of years they worked. It was able to march 54042 employees to their salary and the length of their employment.
* The accuracy of the model is at 0.9662 (96.62%)and the confidence interval at 95% is [0.9647, 0.9677]
* The kappa value of this model stands at 0.507, which show a very good agreement between internal processes/observers. This means that this model is more reliable than the first model based on the difference in the kappa values.
* However, the P-Value is at 2.2e-16. this shows a very significant correlation between employees who earn above R50k and the length of their employment.

## Model conclusion

Based on the kappa value and the P-value, the second model provides is with information that we can rely on. The P-value suggests that there us a significant correlation between the number of years that employees have worked and whether or not they earn above R50 000. The kappa value suggests that should an independent entity review the second model, then we would be able to agree on both our results.

The first model has a very low kappa value, meaning that two different entities would bot be able to agree on the results they come up with, and due to the “Department.Name” being part of the attributes that were chosen for the first decision tree model, the p-value for this model was low. Showing us that the correlation between the employees that earned above R50 000 and the department that they worked for was very low

**R code:**

rdata3<-select(rdataset,Above50k, DepRarity, Agegroup, SalaryType,Years\_Worked )

set.seed(999)

sample3 = sample.split(rdata3$Above50k,SplitRatio = 0.70)

train3 = subset(rdata3, sample==TRUE)

test3 = subset(rdata3,sample==FALSE)

#Training decision tree as a classifier

tree3<- rpart(Above50k~., data = train3)

#predictions

tree.Above50k.predicted3<- predict(tree3,test3,type = 'class')

#evaluating the module with the confusion matrix

confusionMatrix(tree.Above50k.predicted3, test3$Above50k)

prp(tree3)

# Neural Network

## Model Description

* We had to ensure we had normally distributed data for our model to work effectively.
* We had to change our datatypes to allow the machine learning algorithm to learn and perform better.
* We used the “Label Encoding method” to manipulate our raw data before feeding it to the machine learning algorithm.
* We initially used variables that we thought would have a direct impact on whether an employee earns above R50k or not. Namely:
* Age group
* Education
* Occupation
* Months works
* Department Rarity
* We used a deep learning technology called h20, which allowed us to implement the categorical values that we had

## Challenges we had with our model

**Categorical Data**

Most of our data was categorical. This was problematic as neural networks normally take numeric variables as inputs. However we continued with this, so we could compare our results of this model with those of the tree diagram model

**Skewed data**

We realized that the data was relatively skewed towards the above 50k group, this was problematic as it means that employees would generally have a higher chance of earning above 50k. However we continued with this, because we assumed that, that it was intentional to have data that would help has meet our business objective.

## Analysing the results of our original model

From our original model, we had already evaluated and adjusted it according for the most accurate results. The highest accuracy this model could achieve was 72%. This was achieved by:

* Worked from top to bottom to improve the model: by removing one column at a time and then two at a time and so on to see what the success rate would be.
* Working with columns that were binned and categorized
* We got rid of the country column because it was not normally distributed
* This was the highest success rate we found after doing this was 72%, as mentioned above

## Proof of evaluation

The following is how we went about removing columns to try find the greatest accuracy:

**Removing one column**

* Every column without age:71%
* Every column without Education:71%
* Every column without Occuaption:72%
* Every column without Years of residence:68%
* Every column without Department Rarity:53%
* Every column without Years Worked:50%

**Conclusion:**

* Therefore Department Rarity and Years worked are important as they are in the region of a 50% success rate, which makes them the lowest success rate.
* The highest success rate was 72%

**Removing two columns**

* Every column without age and Education:70%
* Every column without Education and Occupation:65%
* Every column without years of residence and age: 72%
* Every column without years of residence and Occupation: 67%
* Every column without age and occupation : 69%

**Conclusion**

* Education and Occupation are more important, as they have the lowest success rate of 65%.
* Years of residence and age are important, as they have the highest success rate at 72%.

**Final Conclusion**

Conclusion : with all tweaks that were made our model stayed the same- at a success rate of 72%.

The final columns that were kept to achieve this rate was:

* Education
* Occupation
* Years of Residence
* Department.rarity
* Months worked
* Salary group

## Assessing the business success criteria

The model we chose was the neural network. In the beginning of this project, we defined our business success criteria as having a model that would meet the following criteria:

* Effectiveness
* Productivity
* Predictive accuracy
* Automated decision making

From the neural network model that we have built, the, the three point of our business criteria was met, namely, effectiveness, predictive accuracy and automated decision making.

## R Code

The R code for the above is included in a separate R script in the folder