**PROJECT 1 MILESTONE REPORT**

**Team No. 19**

(tuple size: 4k, seed-val:35, train:test = 60:40)

Submitted by:

1. Sonika Sarangi Annaiah – 1001733497
2. Vaishakh Deepak Ghati - 1001738753

**PROJECT OVERVIEW**

* The dataset containing the census data consisting of 32,000+ rows was downloaded and imported to R studio as a txt file for analysis.
* The dataset contained over 2000+ missing or unknown(?) values.
* The tuples with missing/unknown values(?) were omitted using the na.omit() function. Over 2,000 rows were omitted and a clean dataset was obtained for analysis.
* The names of variables from the attributes list was taken from the given project description file and a dataframe called ‘df’ was created.
* The seed value was set to 35 as instructed. About 4000 rows were sampled and was split into training data and test data.
* Schema was imposed on the dataset.
* Using the rpart API in R, decision tree was built for the training set. By giving the parameter information gain, the API built the decision tree using information-gain instead of the default gini index.
* The decision tree helped us in understanding the attributes which are used in split criteria.
* The Information.gain API ranked all the attributes in the relative order of Information gain as given below:

attr\_importance

|  |  |
| --- | --- |
| age | 0.086344158 |
| workclass | 0.021142475 |
| fnlwgt | 0.000000000 |
| education | 0.117544148 |
| education\_num | 0.111162798 |
| marital\_status | 0.157030142 |
| occupation | 0.107065152 |
| relationship | 0.172465800 |
| race | 0.007671101 |
| sex | 0.037873947 |
| capital\_gain | 0.097742991 |
| capital\_loss | 0.040897648 |
| hours\_per\_week | 0.059180305 |
| native country | 0.026166586 |

* The priority attributes involved in construction of decision tree were relationship, education and capital\_gain. They were the highest ranked in order of information gain.
* The other attributes that have higher information gain other than the priority attributes was found to be marital\_status, occupation and education\_num.
* After the first branch in rpart API, cp=-1 was used and it developed itself into a complex tree. In a complex tree, occupation and marital\_status were used as attributes other than the priority attributes to perform decision making. (Thereby justifying the utility of the unused priority attributes)
* The predict method was used to predict the test dataset and a confusion matrix was obtained with the predicted values of the test dataset and actual values of the test dataset.

Predicted Values

|  |  |  |
| --- | --- | --- |
| Actual Values | <= 50k | >50k |
| <=50k | 1163 | 171 |
| >50k | 61 | 205 |

The results of Accuracy, Precision, Recall and F1 score was found to be:

Accuarcy : 0.8550000

Precision : 0.8718141

Recall : 0.9501634

F1 score : 0.9093041

If we dropped any attribute that isn’t used in decision making of the tree, we would get the same result. But as a priniciple, we drop the attribute that is of the least importance by information gain, i.e fnlwgt.

Furthermore, education\_num and hours\_per\_week are attributes that are never used in the creation of the complex decision tree or the normal decision tree as well. So these attributes can also be dropped.

Upon withholding these attributes, (since these attributes weren’t used in the building of the decision tree) we get the same result as ones without withholding.

Here the choice of the attributes was based on low ranking of information-gain, and the fact that these attributes were never used in decision making both the simple and complex decision trees.

* The confusion matrix obtained for complex-decision tree is Predicted Values

|  |  |  |
| --- | --- | --- |
| Actual Values | <= 50k | >50k |
| <=50k | 1093 | 157 |
| >50k | 131 | 219 |

The results of Accuracy, Precision, Recall and F1 score was found to be:

Accuarcy : 0.820000 Precision : 0.8744000 Recall : 0.829739 F1 score : 0.8835893

Accuracy,Recall, F-score are lesser than the ones for a non-complex decision tree perhaps due to overfitting.

**FILE NAMES**

1. complete\_project.R – Code where the project is executed.
2. decision\_tree\_final.pdf – Final decision tree which is not complex.
3. decision\_tree\_withcomplexity.pdf – Decision tree with complex consideration (may lead to overfitting)
4. complextree.txt – A text representation of complex tree to understand which attributes are involved.

**CHALLENGES**

* Problems with packages and libraries in R. The libraries used for decision tree modelling were not initially working on R Studio and it was fixed with other supporting libraries before proceeding.
* Cleaning and pre-processing of raw data with missing or unknown(?) values.
* Understanding why some better ranked attributes are not used in building the decision tree.
* Understanding how some values are overfitting.