**IST 707 - Data Analytics**

**Project Report**

**Group 8**

**PREDICTING ADULTS’ SALARIES BASED ON SOCIOECONOMIC ATTRIBUTES**

**Team Members:**

**Jiaming Guo**

**Qi Wang**

**Advait Ramesh Iyer**

**Index:**

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **Topic** | **Pg. No.** |
| 1. | Feature Engineering |  |
| 2. | Unsupervised Learning |  |
| 3. | Supervised Learning |  |
| 4. | Best Model |  |
| 5. | Business Recommendations |  |

**Objective of the Study:**

The core objective of the study is to understand how each socio-economic factor influences salary levels among individuals. To successfully come up with an accurate prediction algorithm, the study involves evaluation and benchmarking of various machine learning algorithms, and their performance the dataset.

**Target Market:**

The report is oriented for Human Resource and Operations professionals, who want to understand how demographics data of employees can be leveraged to gather insights into how salary allocation can be executed, on the basis of education, gender, work experience, as well as financial risk.

The report can also be leveraged by Data Scientists who want to implement deep learning neural networks, with such Machine Learning implementation assembled within one of the hidden layers, with an objective of improving accuracy of predicting salary.

**Methodology:**

The topics that are included in the report are:

1. Feature engineering:
2. Initial Data Exploration
3. Correlation Matrix
4. Unsupervised Learning:
5. Association Rules Mining
6. k-Means Clustering
7. Challenges with Principal Component Analysis and Varimax Rotation
8. Supervised Learning:
9. Decision Tree
10. Naive Bayes Classifier
11. Random Forest
12. K-Nearest Neighbor with Bootstrap
13. Support Vector Machine
14. Logistic Regression
15. Best Model:
16. Linear Kernel for Support Vector Machine
17. Generalized Logistic Regression Model
18. Generalized Logistic Regression Model with Pre-Processing
19. Generalized Logistic Model with Cross-Validation
20. Generalized Logistic Model with Regularization Penalty
21. Business Recommendations:
22. Business Problem
23. Challenges deep-dive
24. Architecture of the Intelligent HR Solution
25. Technical Capability Roadmap
26. Strategic Roadmap

The main business questions that are addressed in the report are:

1. Which socio-economic factors are the most correlated with salary? Are there any specific association rules in the data?
2. Which are the key business metrics that best define the salary structure, and can any specific people strategy improve salary levels?
3. Can the EDA or machine learning based algorithm be generalized in a way to closely simulate the decision-making process similar to an HR executive?
4. Can there be any other attributes in the data-collection process which might improve the predictability of the model?
5. What are the possible ethical issues which will arise by such an HR Analytics solution implementation? How to tackle such a dilemma, and how should the data-ownership value chain be strategized for effective and unbiased implementation?

**1. Feature Engineering**

**1.a) Description:**

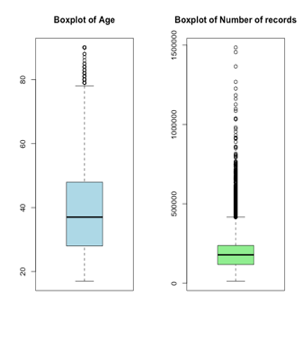
The dataset used for the report consists of 35,000 rows and 15 columns. Following are the attributes in the data:

|  |  |  |
| --- | --- | --- |
| **Variables Name** | **Variable Description** | **Type** |
| **Age** | Record age | Continuous |
| **Workclass** | Record work class 9 levels | Categorical |
| **Fnlwgt** | Number of records for specific kind | Continuous |
| **Education** | Education level | Categorical |
| **Education.num** | Years of education | Continuous |
| **Marital.status** | Marriage status | Categorical |
| **Occupation** | Occupation status | Categorical |
| **Relationship** | Relationship status | Categorical |
| **Race** | Race of the record | Categorical |
| **Sex** | Gender of the record | Categorical |
| **Capital.gain** | Capital gain of the record | Continuous |
| **Capital.loss** | Capital loss of the record | Continuous |
| **Hours.per.week** | Working hours per week | Continuous |
| **Native.country** | Country of resident | Categorical |
| **Salary** | Annual salary, we use 50k as a mid-point | Categorical |

**1.b) Initial Data Exploration:**

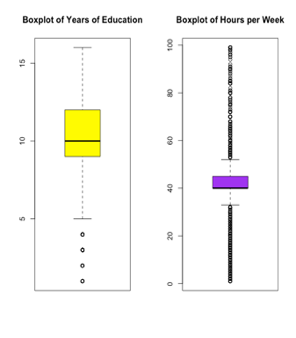
Following are the plots and the conclusions made in the initial data exploration process:

I.



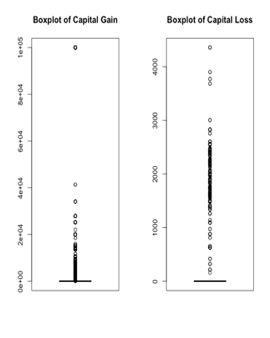
Age and number of records both contain outliers in the upper side.

II.



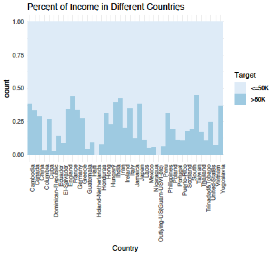
Years of education contains outliers on the lower side, and hours per week has outliers in both the sides of the distribution.

III.



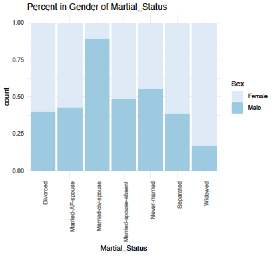
Capital gain and loss both have outliers in the upper side.

IV.



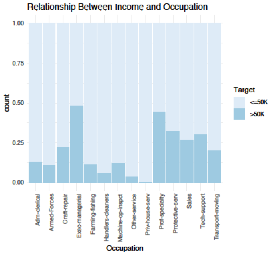
Countries in North America, and Western Europe seem to have higher percentage of employees with salary higher than $50,000.

V.



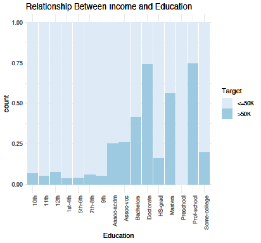
Married and never-married are the biggest categories among Males. Widowed, separated and divorced are the biggest categories among Females.

VI.



Executives, pilots and protective servicemen are the highest paid among all the professions.

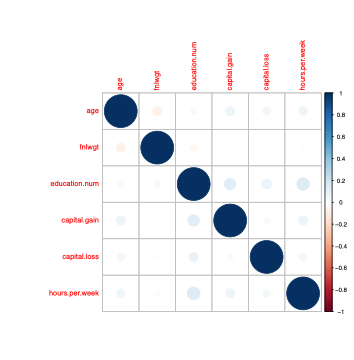
VII.



Doctorate-degree holders, Professors and Master’s degree holders have the highest salaries.

**1.c) Correlation Matrix:**

The correlations among various numeric attributes are as follows:



We considered removing outliers for the correlation matrix, but we did not completely remove outliers in the next steps

Conclusions:

1. Age is negatively correlated with fnlwgt
2. Number of years of education is negatively correlated with fnlwgt
3. Capital gain and loss both are positively correlated with age, years of education and hours per week
4. Hours per week is positively correlated with years of education

**2. Unsupervised Learning**

**2.a) Association Rules Mining:**

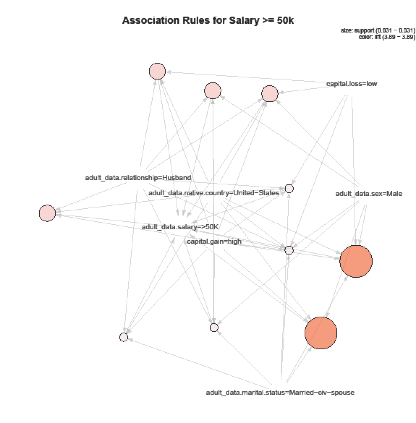
The discretization of the numeric variables was carried out as follows:

1. Age: Teens, Twenties, Thirties, Forties, Fifties, Old
2. Education: Basic, Normal, Advanced, Premium
3. Capital gain and loss: Low, Normal, High
4. Hours per week: Low, Normal, High, Super

In the RHS for the apriori algorithm, “adult\_salary >=50k” was set. Following are the best rules, which were shortlisted first on the basis of lift, and next on the basis of support:

An employee is highly paid if:

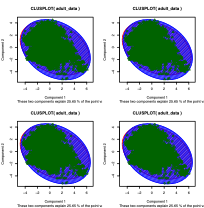
1. The employee is a **male**
2. The employee is a **US citizen**
3. The employee is **married with a spouse**
4. The employee has experienced only a **low capital loss** over his/her career



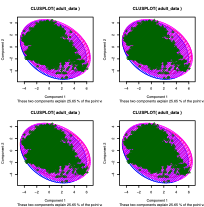
**2.b) K-Means Clustering:**

After implementing Principal Component Analysis (PCA) on the dataset, it could be seen that the first principal component explained 25.65% of the total variability. Following are the outputs of the two k-means iterations:

1. **k=2:**



1. **k=3:**



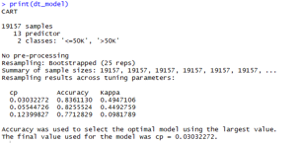
As it can be visualized, k-means is not effective in identifying clusters in the data. Even after conducting Exploratory Factor Analysis (EFA) and Varimax rotation, it is still not able to delineate clusters. Hence, this procedure did not help in any way, and we moved on to unsupervised learning.

**3. Supervised Learning**

**3.a) Decision Tree:**

1. The accuracy of the decision tree model with default setting is **83.53%**
2. The accuracy of the bootstrap decision tree model is **83.61%**
3. The bootstrap performs slightly better
4. According to the tree, following are the features that provide branches with the least possible variance:
5. Married, with spouse
6. Number of years of education
7. High capital gain

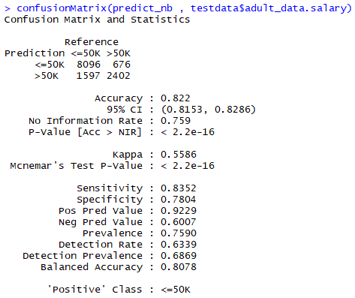
**Output:**



**3.b) Naive Bayes Classifier:**

1. The Naïve Bayes method is used for discrete predictors, which is why the variables were binned before running the algorithm
2. The accuracy of the decision tree model with default setting is **82.2%**

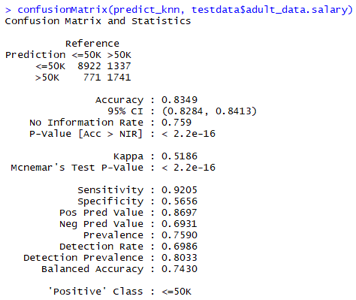
**Output:**

****

**3.c) K-Nearest Neighbor with Bootstrap:**

1. The final kNN model used k=9 to provide the best result
2. The accuracy obtained from the algorithm is **83.5%**

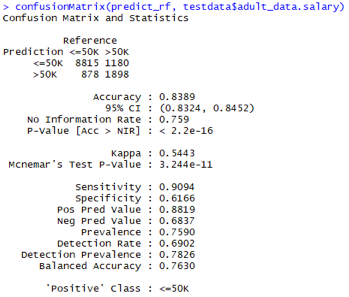
**Output:**



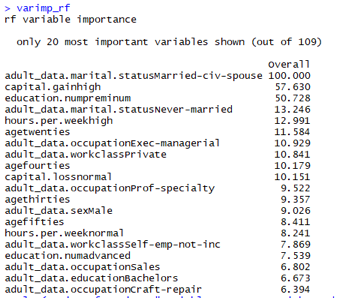
**3.d) Random Forest with Bootstrap:**

1. The model was bootstrapped and ran 25 times
2. The model provided an accuracy of 83.9%

**Output:**



**The variable importance for the top-20 variables are as follows:**

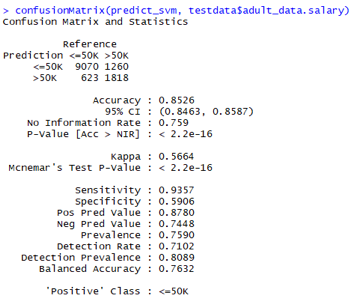
****

The variables with high importance can further be used for assigning weights to the predictions using a penalty method, and then the ML model can be evaluated accordingly. However, we are still yet to conclude which shallow learning model is the best, and plan to use this knowledge in that model.

**3.e) Support Vector Machine with Bootstrap:**

1. The training data was resampled 25 times within the bootstrapping process
2. The accuracy of the SVM model was 85.26%

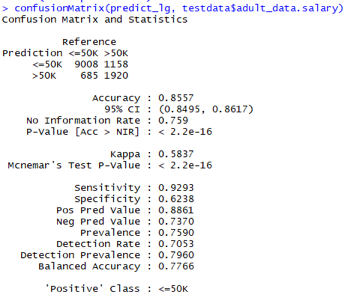
**Output:**



**3.f) Logistic Regression with Bootstrap:**

1. The training data was resampled 25 times within the bootstrapping process
2. The accuracy of the Logistic model was 85.38%

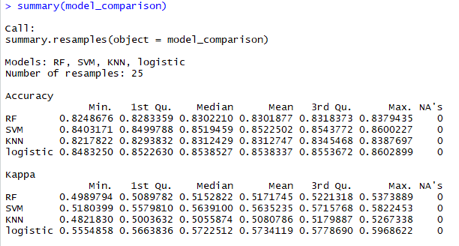
**Output:**

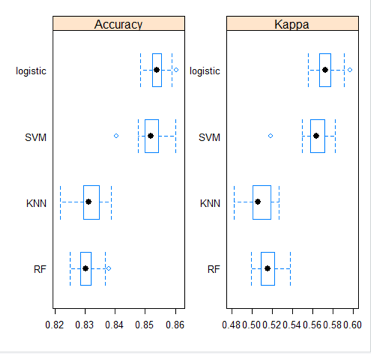


**Model Comparison:**

As per the comparison of all the above models, it can be concluded that the Logistic Regression model with Bootstrapping performs better.

**Output:**





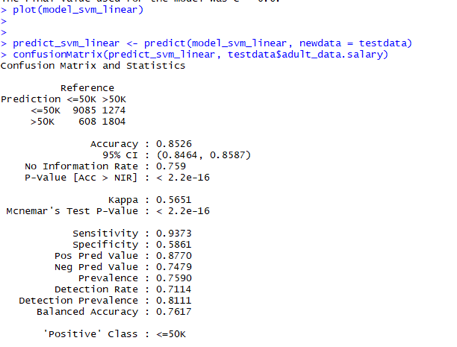
**4. Best Model Selection**

We shortlisted Support Vector Machine and Logistic Regression for further evaluation, because they were the best shallow learning models. We started with Support Vector Machine model.

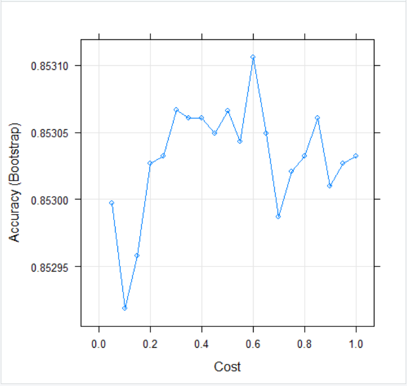
**4.a) Linear Kernel for Support Vector Machine:**

1. Parameters setting:
   1. Method: svmLinear
   2. Pre-processing: centered and scaled
   3. Bootstrapped with 25 repetitions
   4. Cost range (0,1,0.1)
2. Achieved an accuracy of 85.26%, which is similar to the accuracy through radial-basis function

**Output:**

****

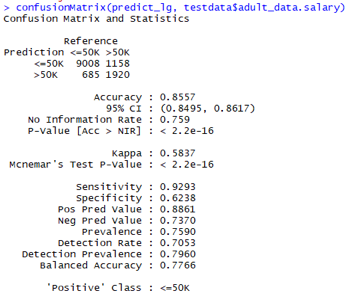
The best SVM model provides an accuracy just above 85.3%, which means we will not be able an accuracy equivalent to Logistic model.

****

**4.b) Generalized Logistic Regression Model:**

1. Parameters setting:
   1. aMethod: glm
   2. Pre-processing: No
   3. Bootstrapped with 25 repetitions
2. Achieved an accuracy of 85.57%, which is much better than simple logistic regression model
3. Due to the promising gain in accuracy, we decided to investigate the model further

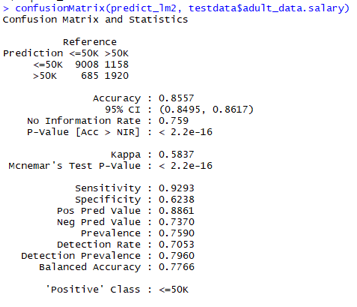
**Output:**

****

**4.c) Generalized Logistic Regression Model with Pre-processing:**

1. Parameters setting:
2. Method: glm
   1. Pre-processing: centered and scaled
   2. Bootstrapped with 25 repetitions
3. Achieved an accuracy of 85.57%, which is the same as the GLM model

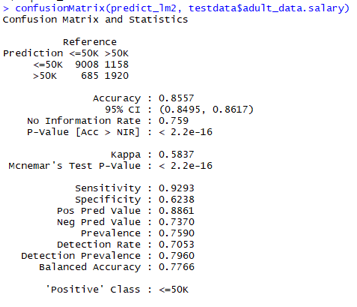
**Output:**



**4.d) Generalized Logistic Regression Model with Cross-validation:**

1. Parameters setting:
2. Method: glm
   1. Pre-processing: centered and scaled
   2. Cross-validation: 10-fold
   3. Bootstrapped with 25 repetitions
3. Achieved an accuracy of 85.57%, which is the same as the GLM and pre-processed GLM model

**Output:**

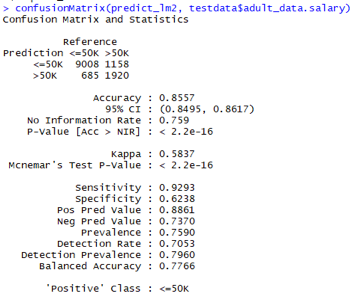
****

**4.e) Generalized Logistic Regression Model with Regularization Penalty:**

1. Parameters setting:
2. Method: glmnet
   1. Pre-processing: centered and scaled
   2. Cross-validation: 10-fold
   3. Bootstrapped with 25 repetitions
   4. Alpha range(0.1,1,0.1)
   5. Lambda range(0.001,1,0.111)

2. Achieved an accuracy of 85.61%, which is the best accuracy out of all the models

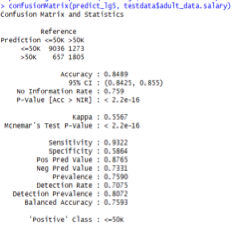
**Output:**



**4.f) Generalized Logistic Regression Model with Regularization Penalty:**

**Association Rules Mining:**

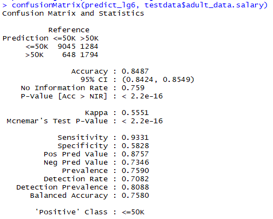
1. Regularization Penalty (Alpha (0.1,1,0.1) lambda(0.001,1,0.111)
2. Selected input: capital.gain+capital.loss+adult\_data.education+adult\_data.relationship+hours.per.week+adult\_data.occupation+adult\_data.sex+adult\_data.marital.status
3. The accuracy of the model decreases



**4.g) Generalized Logistic Regression Model with Regularization Penalty:**

**Identified top rules:**

1. Regularization Penalty (Alpha (0.1,1,0.1) lambda(0.001,1,0.111)
2. Selected inputcapital.gain+ capital.loss+adult\_data.education+adult\_data.relationship+hours.per.week+adult\_data.occupation+adult\_data.sex+adult\_data.marital.status
3. The accuracy of the model decreases



**Conclusion:**

The Generalized Logistic Regression model with Regularization Penalty used on all attributes, performs the best.

**5. Business Recommendations**

**5.a) Business Problem:**

**Key Problems:**

1. Salary mapping in accordance with employee portfolio
2. Dynamic higher-management compensation adjustment
3. Record management for socio-economic attributes and performance KPIs

**Main Context behind the problems:**

1. Retention/attrition rate
2. Compensation allocation
3. Past/present employee history
4. ERP integration
5. CRM Integration
6. Predictive capabilities in the Human Resource solution

**Main issues hindering the intelligent solution:**

1. Relational databases not inter-connected
2. Missing data-points
3. Incorrect data-points considered in the system, leading to wrong predictions
4. In-efficient resource allocation
5. Customer journey with Sales/Marketing team not mapped and modeled

**5.b) Challenges Deep-dive:**

**Challenge 1: Garbage Data**

1. Collecting the wrong data points for developing predictive analytics capability, leading to incorrect predictions
2. Data is sparse in nature, which leads to overfitting/underfitting, or inability to run certain algorithms
3. Data is incorrectly entered, due to lack of auditing
4. Data points not mapped, leading to inability of reaching full predictive potential

**Challenge 2: Lack of Technological/Managerial Capability**

1. Lack of cloud-infrastructure for big-data analytics
2. Lack of design and reasoning skills when creating survey
3. Inability to engage employees in filling out surveys
4. Internal politics in the organization leads to generation of wrong or sparse data points, due to favouritism
5. Lack of knowledge in cross-functional management practices and methods

**Challenge 3: Unclear Plan of Action**

1. Lack of direction and approach, whether bottom-up or top-down
2. Inability to gather requirements of the system
3. Lack of KPI metric identification and evaluation strategies
4. Inability to identify which method is applied in which stage of implementation
5. High-dimensionality causes interpretability problems
6. Overfitting/underfitting problems

**5.c) Implementation:**

**Zero - year Milestone:**

**Technical standpoint:**

Data Procurement and engineering: The following steps can be implemented in the 0-year milestone

1. Collecting data
2. Data cleaning and filtering
3. Noise reduction
4. Correlation Tests
5. Heteroscedasticity
6. PCA/clustering
7. Statistical significance tests

Data Storage:

1. Unstructured - images, text, social media data
2. Structured - JSON, XML
3. Scaling them via Spark, Hadoop, MapReduce, Cassandra and NoSQL
4. Improving processing capabilities through cloud integration

**Strategic standpoint:** Hire data scientists, engineers, business analysts with domain expertise

**Two - year Milestone:**

**Technical standpoint:**

Traditional Machine learning:

1. Linear/Logistic regression
2. SVM
3. KNN
4. Naive Bayes
5. Decision Trees

Sampling and Ensemble Learning:

1. Sampling tests
2. Bootstrap aggregating
3. Boosting
4. Bucket of Models
5. Stacking
6. ROC curve and Kappa metric evaluation

**Strategic standpoint:** Identify operational, strategic and employee metrics. Introduce capabilities for descriptive and predictive analytics, as well as ML/AI capabilities.

**Four - year Milestone:**

**Technical standpoint:**

Knowledge infrastructure and pricing strategy:

1. Knowledge centre development
2. Content strategy for thought leadership
3. CRM integration strategy
4. ERP integration strategy
5. Salary-allocation strategy

Artificial intelligence capabilities:

1. Recurrent Neural Networks for time-series data
2. Convolutional Neural Networks for image recognition
3. Deep neural networks for custom requirement
4. Semi-supervised learning capabilities
5. Evolutionary machine learning methods for optimization
6. Natural language capabilities through Deep NLP networks

**Strategic standpoint:** Scale the big-data infrastructure to integrate all CRM, ERP and HR Intelligence solution. Integrate and optimize the systems, and inculcate salary prediction methods within the intelligent solution.

**Conclusion:**

The report provides a comprehensive insight about how the dataset can be analyzed in the best possible way, as well as provides a phase-wise implementation roadmap for the Intelligent Human Resource Solution.