Part I Report -Executive Summary

Executive Summary

https://github.com/huangjjv1/MISCADA_CoreIIA_Classification

1) Data introduction & Objectives

We are researching what factors are most relevant to people's salary. We can finally know through the results which aspects of a person should be improved or should be valued to increase their income. This analysis can be submitted to the government as a basis for guiding public spending, while increasing social productivity. Similarly, if we have the features of a person, then we can predict the approximate range of his/her salary expectation.

Through the research, we have a set of data containing about 32500 personal samples. Each sample is about the information of persons in 15 kinds of features including Age, WorkClass, numerical id, Education (degree) 5. EducationNumber (years a person has been educated) 6. Martial Status 7. Occupation 8. Relationship 9. Race 10. Gender 11. CapitalGain 12. CapitalLoss 13. HoursWeek 14. NativeCountry 15. Income.

The problem is binary classification because we only focus on whether a person's annual salary reach 50k/year.

2) Modelling & Implement Strategy

We now have proposed 3 methods of analysis:

1. **deep learning:** It has more flexibility to process a large amount of data. However, it calls for high computer hardware requirements. That makes it an algorithm with the highest upper limit and the highest cost.

2. **KNN**: Does not require training costs and has high stability of output in accuracy, but is not compatible with all different data.

3. **random forest**: It is possible to extract the importance of each feature through certain feature analysis, which has strong interpretability. Only a small amount of data preparation is needed to process both categorical and numerical data. Besides it also has good stability and reliability.

3) Result Analysis & Interpretable part

Through Initial data exploration, we can get the correlation between every two dominant features. We can observe their common changes, trends, whether there is a common increase or decrease. According to our intuitive experience, wages should be related to the industry in which they are located, that is, the type of work. And the level of work, which can be consider as the working age. According to correlation analysis we find that, the higher the education numbers, working age, working hours per week, the higher the salary. Married persons and those are in executive positions earn more than those are not, which is quite interesting.

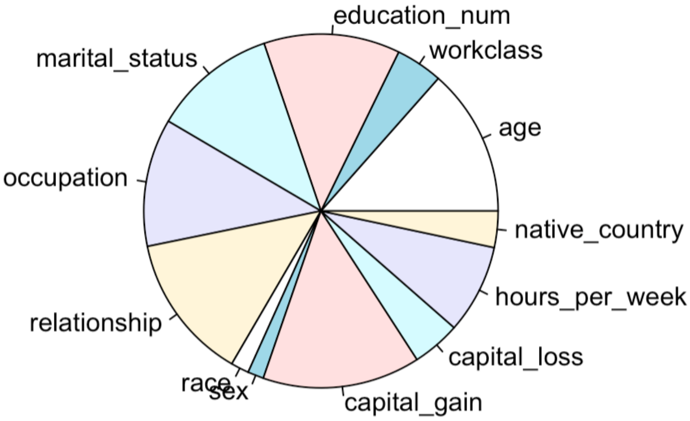


Fig. 1 Importance of each feature

From the figure above we can conclude that age, years of education time, marital status, occupation, relationship, capital gain and work hours per week are the more important features of persons that can decide their salary levels. It is worth mentioning that gender and race are not very distinct from other factors. It also helps reduce racial and gender discrimination in workplace culture.

The importance information we got here is just the target we want. However, we can still do the predictions by training models to discover the potential of the data.

4) Predictions

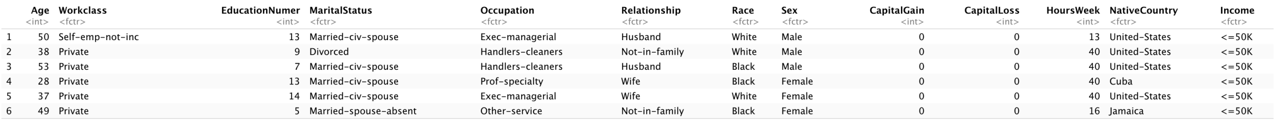
Through a series of data processing, we have extracted the main features. We now can train the models using these observations with those several features. After training models with machine learning, cart, random forest and KNN algorithms, we finally confirm the random forest is the ideal model we want. The stability and understandability of the random forest in the selection of the model play an outstanding role in this analysis.

The accuracy rate we obtained through prediction can exceed 85%. The stability and understandability of the random forest plays an outstanding role in this analysis. We have not only analyzed the important factors, but also made predictions on the test samples provided. That is to say, we have achieved and met the expected results. It can be expected that the analysis of this data will play a key role in future human resources and social issues

Part II Report - Technical Summary

Technical Summary

1) Initial data exploration



	Age	Workclass	EducationNumer	MaritalStatus	Occupation	Relationship	Race	Sex	CapitalGain	CapitalLoss	HoursWeek	NativeCountry	Income
	<int>	<fctr>	<int>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<int>	<int>	<int>	<fctr>	<fctr>
1	50	Self-emp-not-inc	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
5	37	Private	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
6	49	Private	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K

Fig. 1 Data (Removed col3 &4) (accessible on web: https://archive.ics.uci.edu/ml/datasets/Adult)

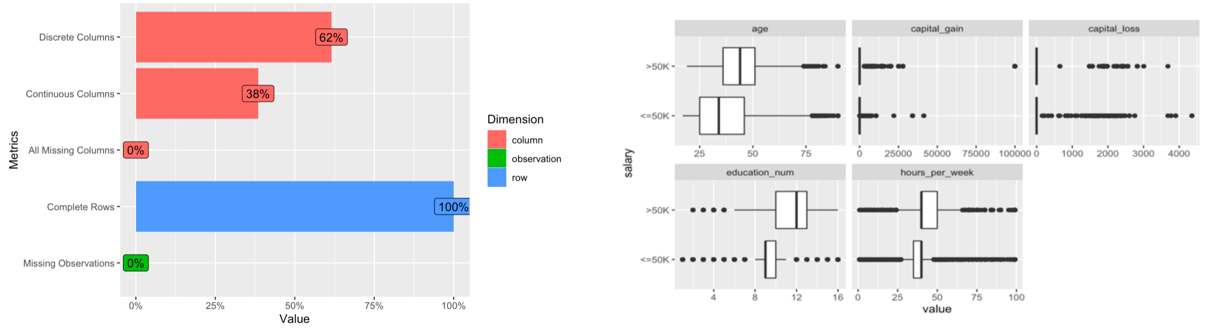


Fig. 2 data feature

The data is about the information of persons in 15 dimensions with 38% of columns are described in continuous value. Among features with continuous values, age, education_num and hours_per_week have relatively higher differences than capital_gain/loss.

Prediction task is to determine whether a person makes over 50K a year which is the 13rd feature of the data. The model is a binary classification. To achieve the goal, I have tried to implement three methods: deep learning (Main method), KNN and random forest algorithms to complete the classification of people's annual salary. And then, we can do the prediction by the models and analyse performance results.

2) Details of any data coding or feature engineering

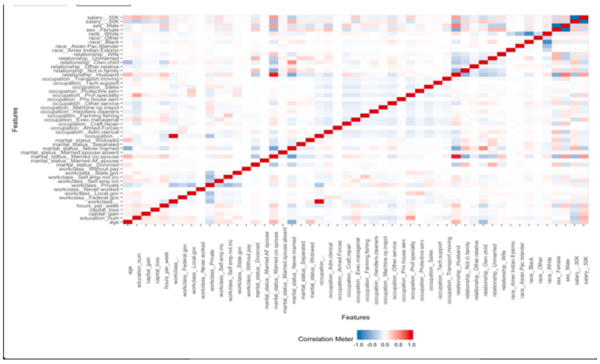
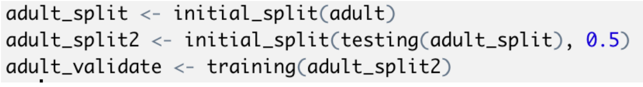


Fig. 3 Correlations among features

Column names: The adult data has 32560 observations and 15 features per observation. Features are respectively: 1. Age 2. WorkClass 3. Final weight (a special id of each observation) 4. Education (degree) 5. EducationNumber (years a person has been educated) 6. MartialStatus 7. Occupation 8. Relationship 9. Race 10.Sex 11. CapitalGain 12. CapitalLoss 13. HoursWeek 14. NativeCountry 15. Income.

Feature engineering: Three features of variables have suspicious high correlations. We can consider the id (feature 3) of each observation is a random value, so we can remove it from the columns; The education years can be considered to show the similar meaning of education degree, so we also remove the education_nums from the columns. Thus, we have 12 features working as variables, and salary conditions as response.

Train/test/validate splitting



```
adult_split <- initial_split(adult)
adult_split2 <- initial_split(testing(adult_split), 0.5)
adult_validate <- training(adult_split2)
```

The training set is used to train the model parameters, the test set is used to estimate the generalization error of the model on the sample, and the validation set is used to get the fit hyperparameters of the model. In this case, validation set is from the other half of the raw data. Test set has already been given. So, we just use it.

Strategies used to address any missingness

Although the figure in the first section shows that the NA rate is 0, it is actually not 0 but very close to 0. In addition, NA only emerges in non-continuous features. That is to say, the ratio is small enough that we can give NA a specific given value. In Deep Learning model, the strategy here we implement is to set them to unknown feature, and this will not affect too much on the result according repeated tests. While in Random Forest model we use Median of corresponding column (na.action = na.roughfix).

3) The approach taken to model fitting, including any model design, early stopping criteria, hyperparameter selection or tuning, and algorithm choices

To be able to analyse data at multiple levels, we carry out KNN, random forest, and deep learning algorithms to do the analysis. **Deep Learning**: 1) As we know that if the layers are not configured properly (especially when setting up the network without dropout layer), overfitting may appear. To tune the hyperparameter, we can observe that after about 20-30 epochs, the val_loss get higher obviously or stable. So, I take the best performance (generalization error) during early stopping epoch;

Random Forest: I mainly tune the number of variables randomly sampled as candidates at each split. After several tests the trend emerges, The mtry best is 3. Then, the error rate (using OOB estimate) can be the lowest: 0.18.

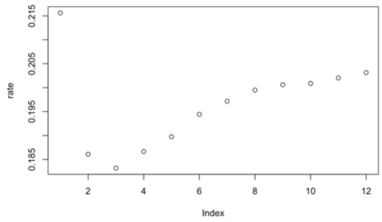


Fig. 4 Random forest error

KNN: Tune the k value, and it turns out to be 1 with auc value 0.9944 (This value is doubtfully high). We mainly discuss Deep Learning and Random Forest method.

4) Insights into improvements achieved through different architectures (deep learning)

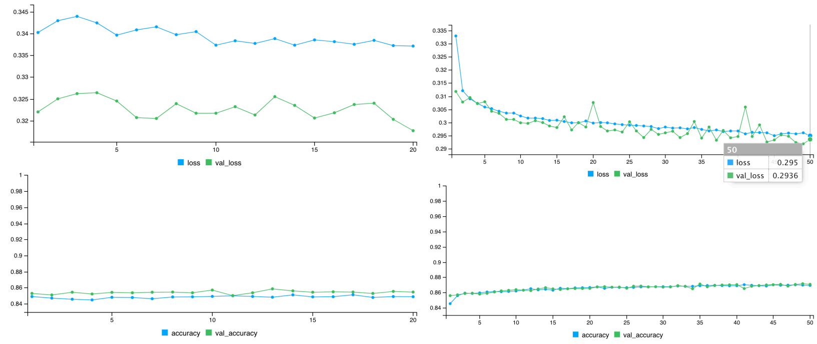
Machine Learning architectures: A basic of combination of layers consists of one density layer (using relu activation function to complete a non-linear function transformation on the input data), one normalization layer and a dropout layer. The first one is where parameter and corresponding loss function values are at; The second one is to normalize its activation value to a normal distribution with a mean of 0 and a variance of 1. The last one is the so-called dropout layer. It let the activation value of a certain neuron stop working with a certain probability p, which can make the model more general, because it will not rely too much on some local features.

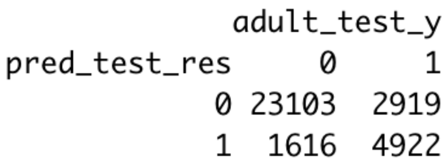
Especially if the model has too many parameters and too few training samples, the trained model is prone to overfitting. It is worth mentioning that DNN is trained batch by batch. The first reason is to better deal with non-convex loss functions. In the case of non-convex, even if the full sample is calculated, it will be stuck on the local optimization. The batch indicates the partial sampling of the full sample, which is equivalent to artificially introducing correction sampling noise on the gradient. That makes it more likely that all the way to find another way to search for the optimal value; the second is to reasonably use the memory capacity. So, when we nest multiple combination of these layers to the network, we simultaneously adjust the batch size to increase the calculation efficiency and the fitting performance.

5) **Details on the performance of the model, including calibration. Reporting of loss function choices, any post-model analysis such as tuning true/false positive rates, and justification of alternative objective functions**

We need calibration when doing online model evaluation. We will decide whether to do calibration based on predict data and real feedback.

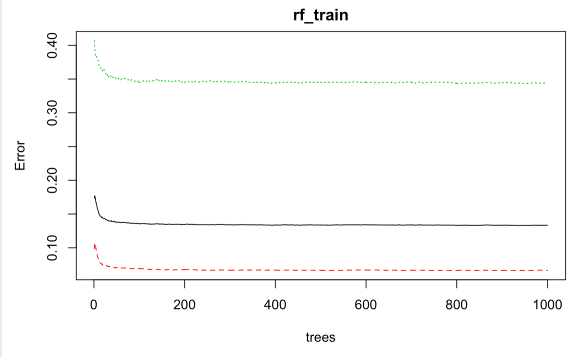
Performance of Deep Learning and Random Forest:

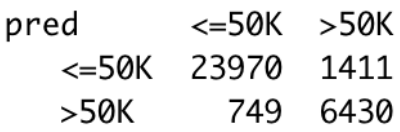




```
             adult_test_y
pred_test_res     0     1
            0 23103  2919
            1  1616  4922
```

Fig. 5 simple DNN. Tuned DNN and confusion matrix of DNN





```
pred      <=50K  >50K
   <=50K  23970  1411
   >50K     749  6430
```

Fig. 6 Performance plot and confusion matrix of RandomForest

Area under the curve: 0.7812 (DNN) and 0.8949 (RandomForest).

DNN: Loss function is set to be binary cross-entropy because it is a binary classification; optimizer is set to Rmsprop, which tries to solve the problem that the gradient may vary greatly in amplitude. Some gradients may be small, while others may be large, which can cause a very tricky problem-trying to find a single overall learning rate for the algorithm. If we use full batch learning, we can use only gradient symbols to solve this problem. In this way, we can guarantee that the size of the weight update is the same. This adjustment helps to greatly solve the saddle point and stationary problems, and we can take large enough steps even with small gradients. Metrics uses accuracy.

6) Any efforts at interpretability

Considering that this is not a network like convolutional neural network, it is difficult to perform interpretable processing in neural networks. The actual meaning of the node parameters is difficult to have an intuitive reflection and dimensionality reduction processing. So, in order to give non-scientific background readers an intuitive understanding of the model, through the establishment of a random forest model and the use of the Gini algorithm, we can get the weight (importance) of each feature

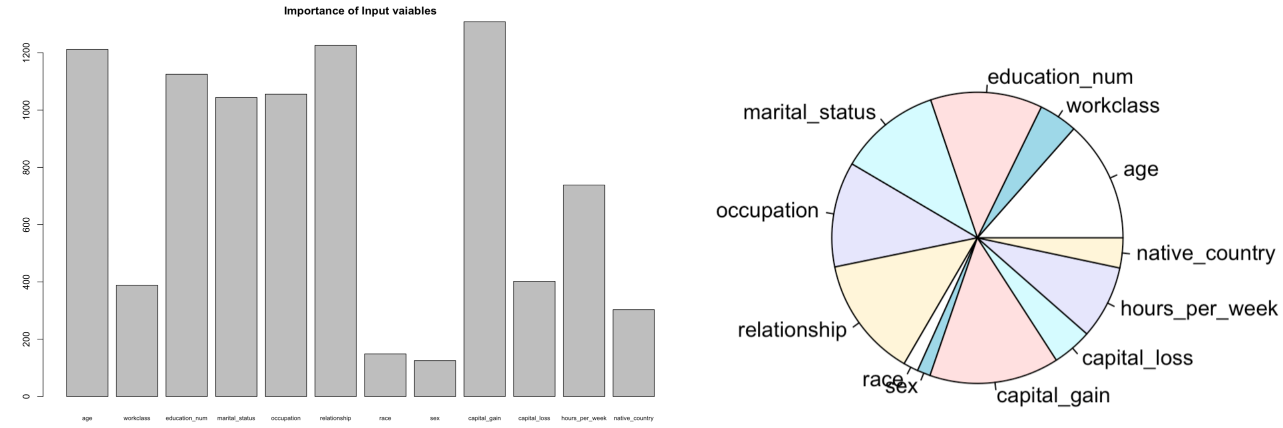


Fig. 7 Importance of each feature

From the figure above we can conclude that age, years of education time, marital status, occupation, relationship, capital gain and work hours per week are the more important features of persons that can decide their salary levels.

7) Conclusion

Overall, random forest is more accurate and more interpretable for binary classification models, which can be reflected from the auc value. Deep learning has the disadvantages of training models, optimizing models, and resource consumption. However, it may have more advantages in the establishment of more multi-dimensional and models and general optimization. In general, for this binary classification problem, random forest is the more appropriate model to choose.