# PREDICTING ADULTS' SALARIES BASED ON SOCIOECONOMIC ATTRIBUTES

Data Analytics (IST 707)
Group 8
Final Presentation

Team Members: Advait Ramesh Iyer, Qi Wang, Jiaming Guo

| 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 |

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.

Proposal of an intelligence Human Resource solution

# Business Questions

Which socio-economic factors are the most correlated with salary? Are there any specific association rules in the data?

Which are the key business metrics that best define the salary structure, and can any specific people strategy improve salary levels?

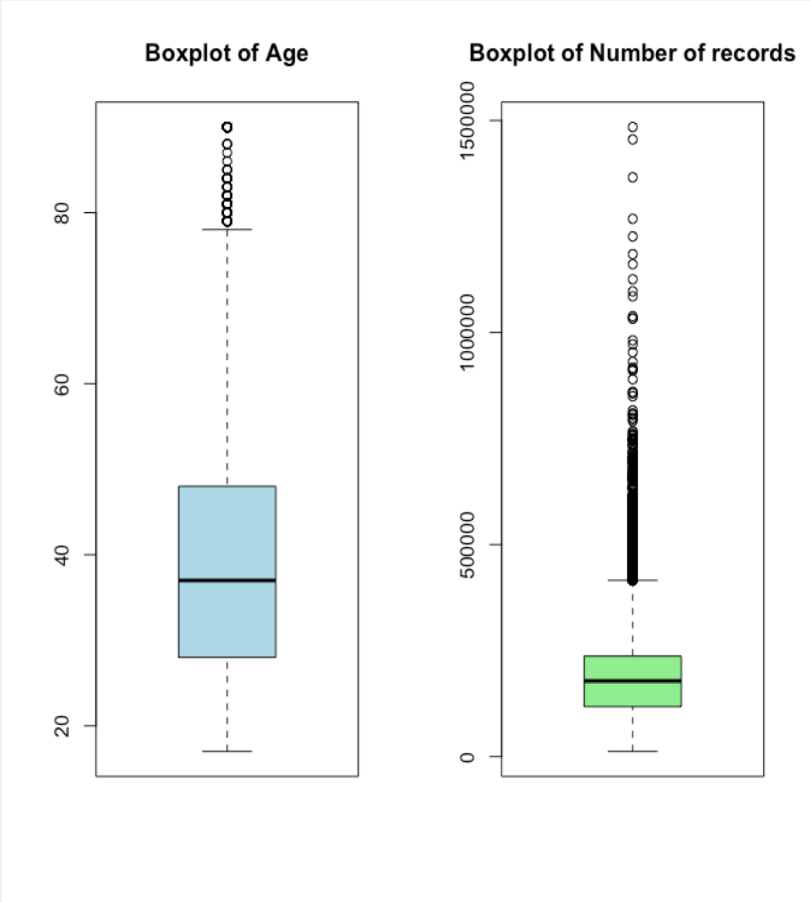
Can the EDA or machine learning based algorithm be generalized in a way to closely simulate the decision-making process similar to an HR executive?

Can there be any other attributes in the data-collection process which might improve the predictability of the model?

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?

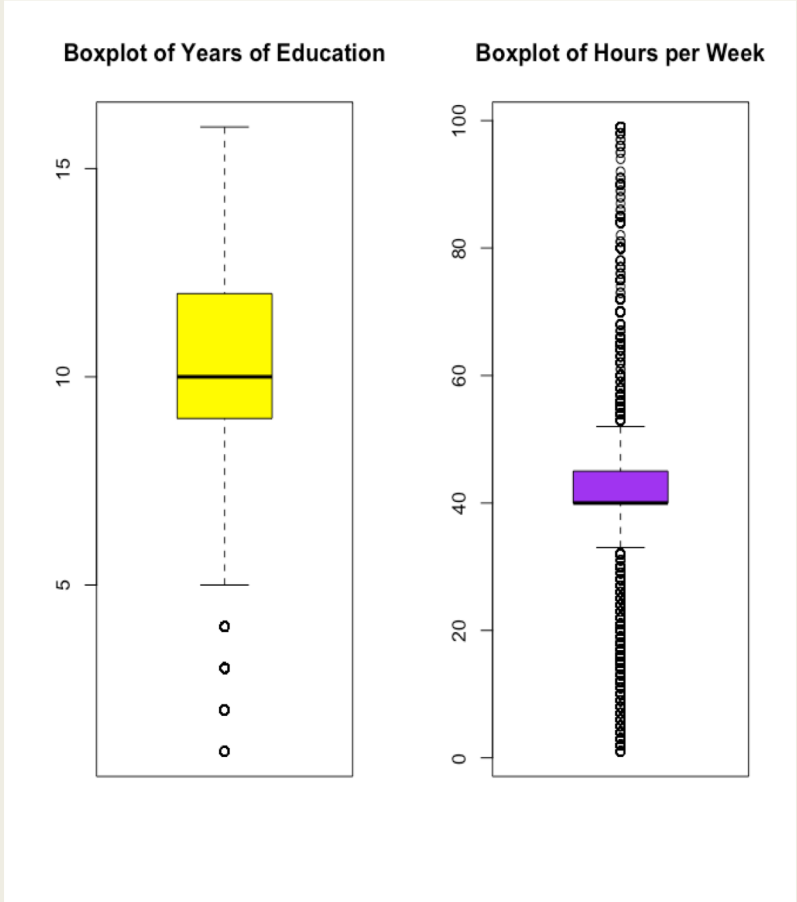
# INITIAL DATA EXPLORATION

# Data Exploration: Continuous Variables



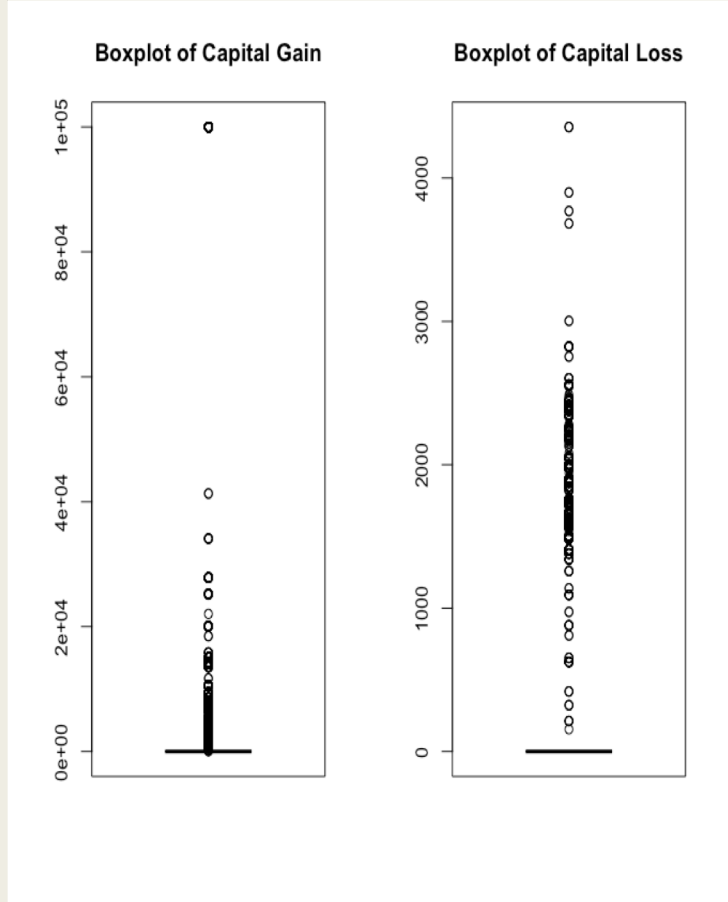




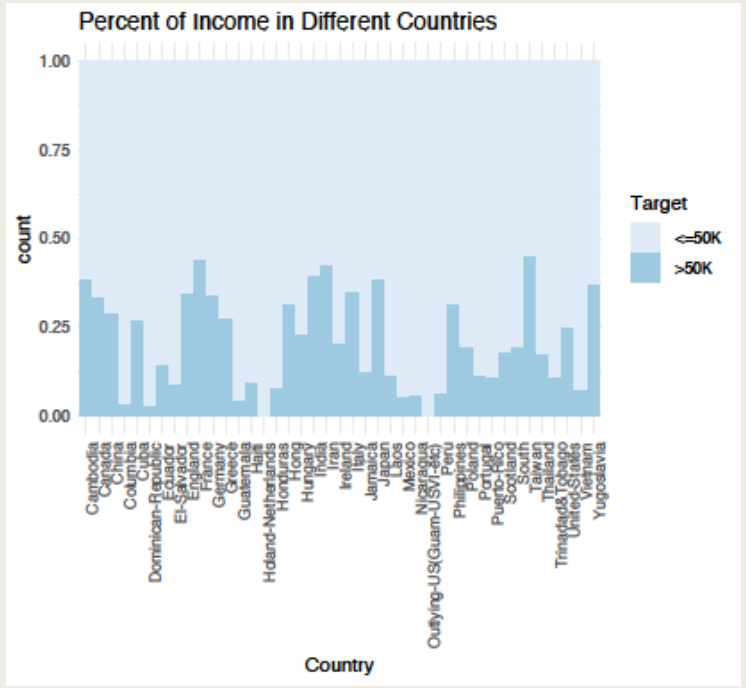


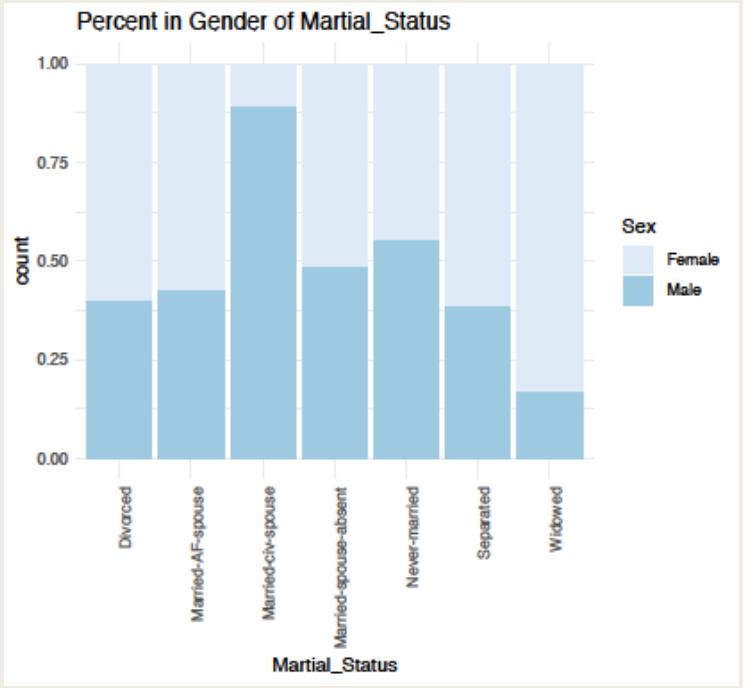


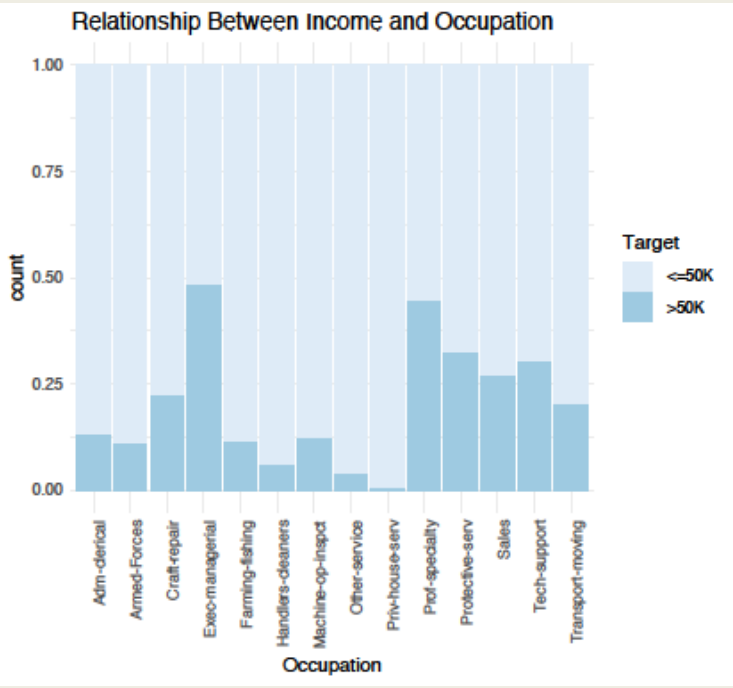




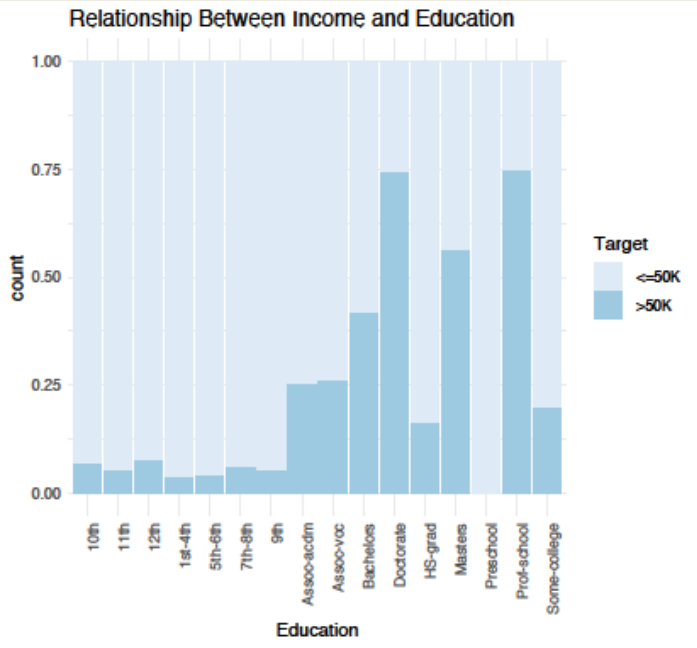
# Data Exploration: Categorical/Multi-dimensional plots (1)

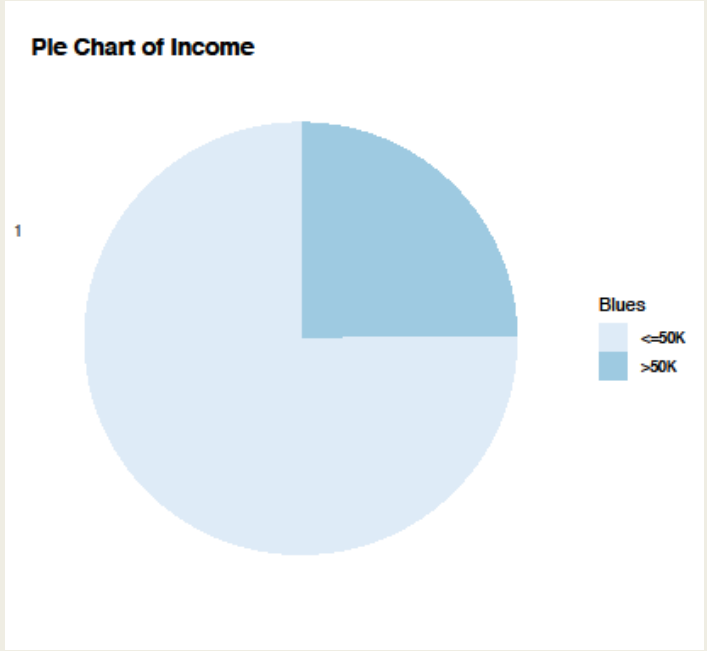


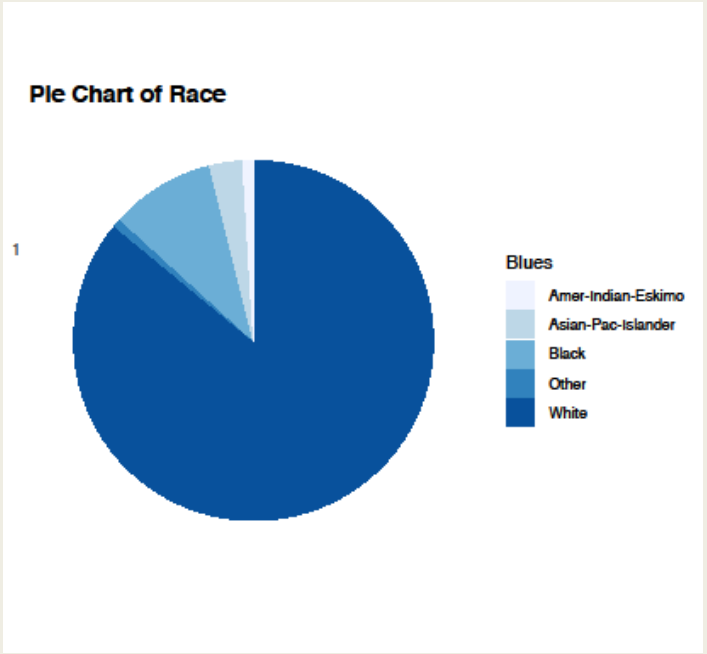



# Data Exploration: Categorical/Multi-dimensional plots (1)

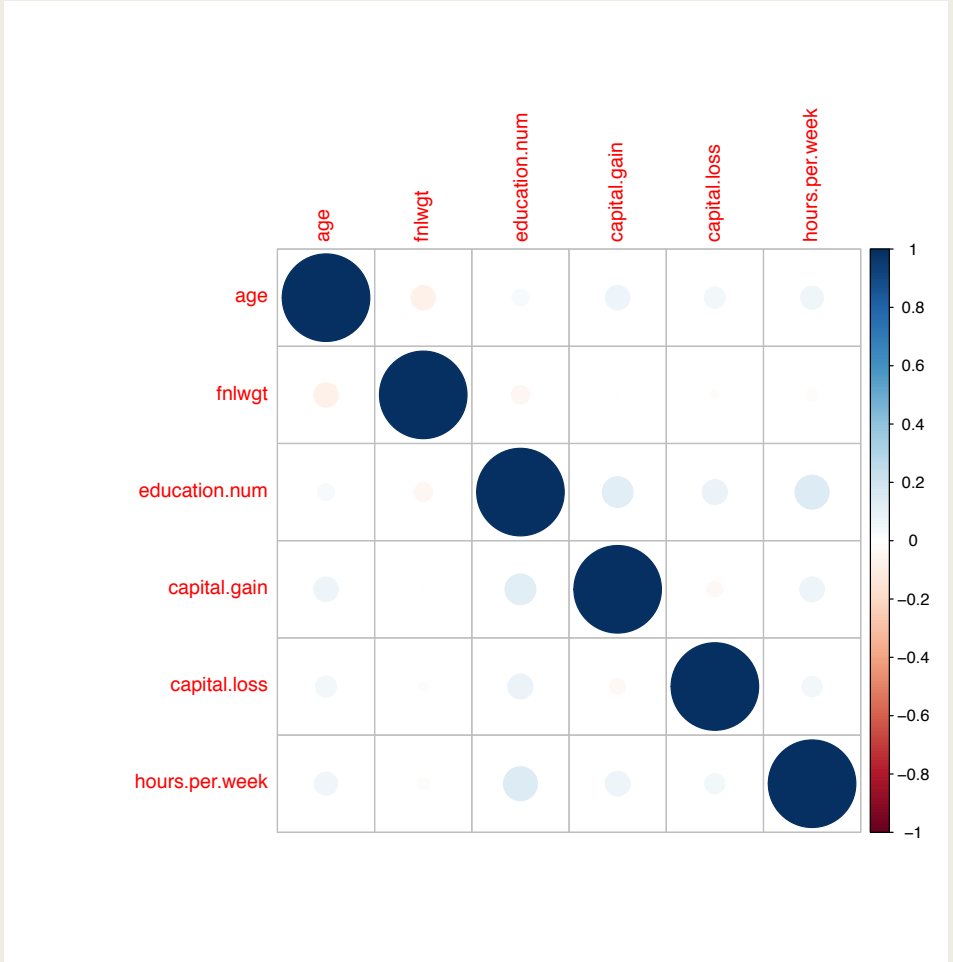


# Correlation Matrix: Continuous Variables

- We considered removing outliers for the correlation matrix, but we did not completely remove outliers in the next steps
- Age is negatively correlated with fnlwgt
- Number of years of education is negatively correlated with fnlwgt
- Capital gain and loss both are positively correlated with age, years of education and hours per week
- Hours per week is positively correlated with years of education

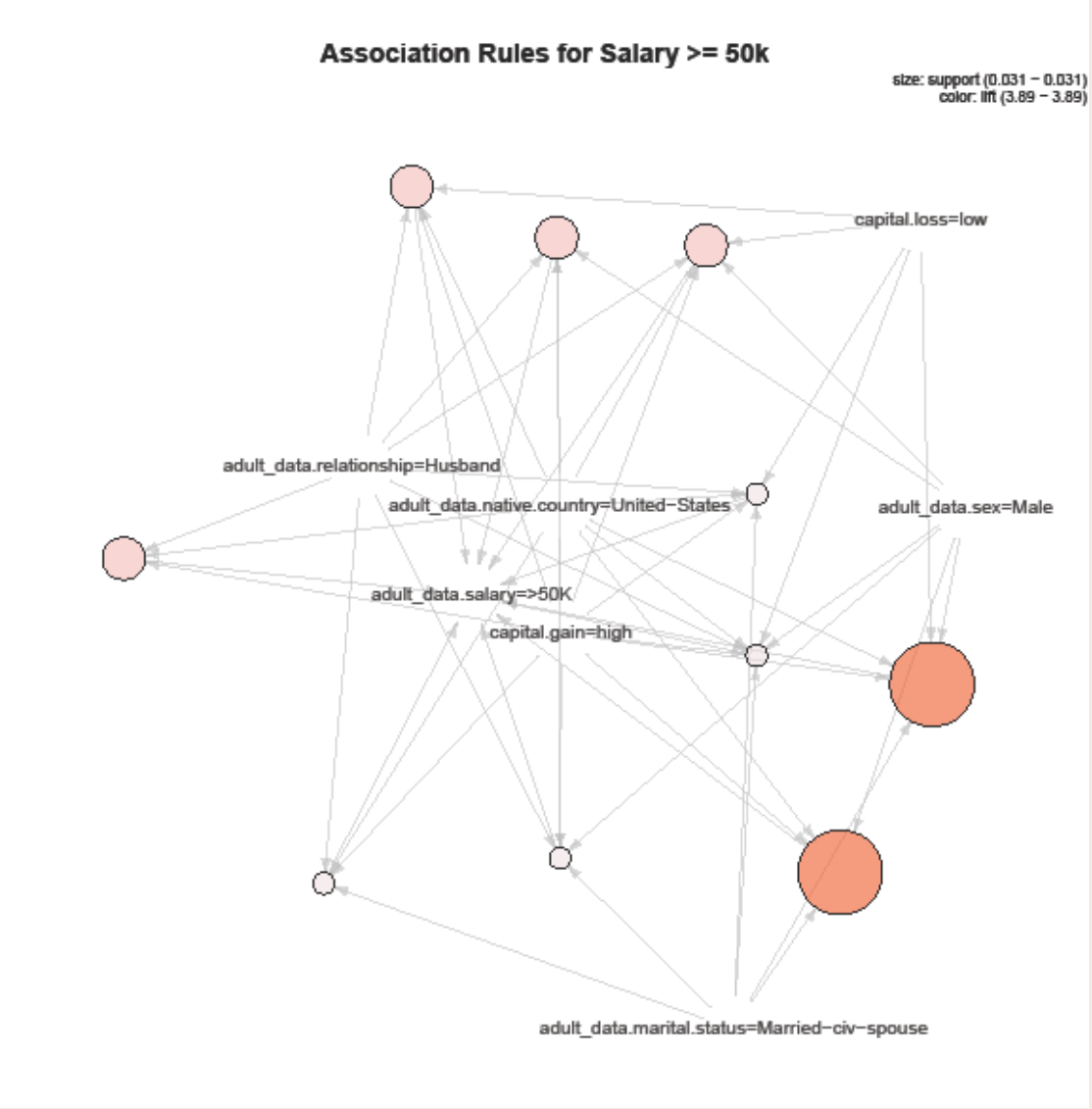


# UNSUPERVISED LEARNING

# Association Rules Mining

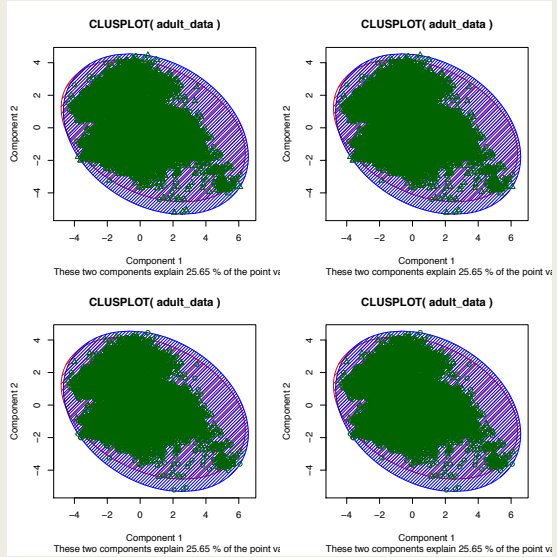
Discretized:

- Age: Teens, Twenties, Thirties, Forties, Fifties, Old
- Education: Basic, Normal, Advanced, Premium
- Capital gain and loss: Low, Normal, High
- Hours per week: Low, Normal, High, Super
- A person is highly paid if:
  - *Male*
  - *United States citizen*
  - *Married with spouse*
  - *Low capital loss*

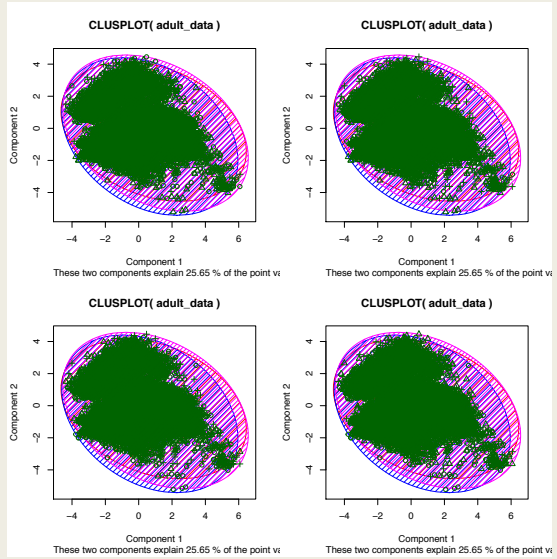


# k-Means Clustering

- The top two principal components (PCs) explain 25.65% of total variability
- The centroids for clusters are not really far, so there are no valuable inputs that we got through k-Means clustering
- This means, there is not much variability in the data, which could have led to clear segmentation
- Varimax rotation fails to optimize centroids in a way which makes sense




2 Clusters




3 Clusters

# SUPERVISED LEARNING

Supervised Learning Models

| Algorithm | Accuracy |
| --- | --- |
| Decision Tree | 83.5%, Bootstrap: 83.6% |
| Naïve Bayes Classifier | 82.2% |
| Random Forest | Bootstrap: 83.5% |
| K-Nearest Neighbour | Bootstrap: 83.9% |
| Support Vector Machine | Bootstrap: 85.3% |
| Logistic Regression | Bootstrap: 85.4% |

We choose SVM and Logistic Regression for further evaluation,

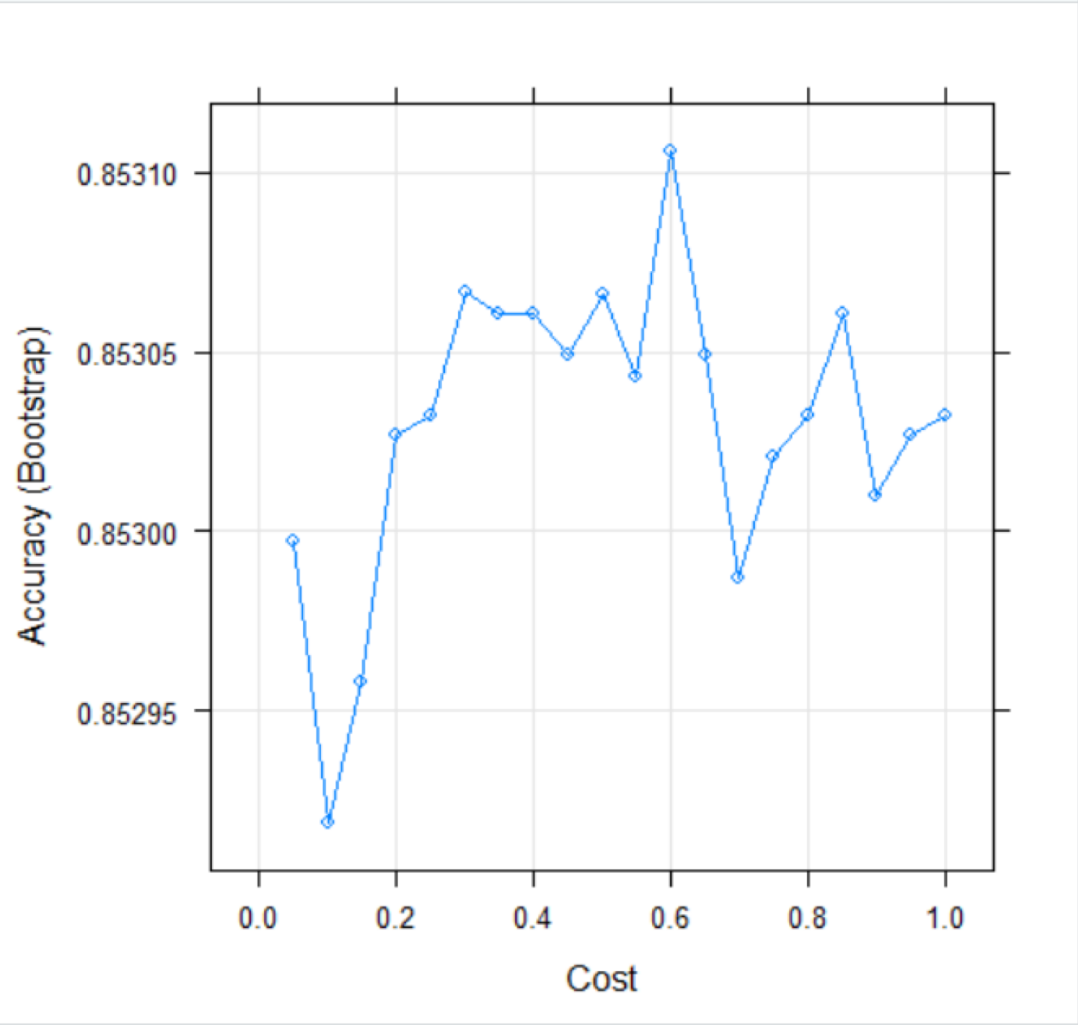
# Linear kernel for Support Vector Machine

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

```
> plot(model_svm_linear)
>
>
> predict_svm_linear <- predict(model_svm_linear, newdata = testdata)
> confusionMatrix(predict_svm_linear, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9085 1274
     >50K    608 1804

               Accuracy : 0.8526
                 95% CI : (0.8464, 0.8587)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5651
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9373
            Specificity : 0.5861
         Pos Pred Value : 0.8770
         Neg Pred Value : 0.7479
             Prevalence : 0.7590
         Detection Rate : 0.7114
   Detection Prevalence : 0.8111
      Balanced Accuracy : 0.7617

       'Positive' Class : <=50K
```

# Linear kernel for Support Vector Machine

- According to the plot between accuracy through bootstrap, and the cost function, it was identified that the highest possible value of accuracy was just above 85.31%, which is not good enough
- Considered dropping the model, and focusing on generalized Logistic Regression model



# Generalized Logistic Regression Model

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

```
> confusionMatrix(predict_lg, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9008 1158
     >50K    685 1920

               Accuracy : 0.8557
                 95% CI : (0.8495, 0.8617)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5837
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9293
            Specificity : 0.6238
         Pos Pred Value : 0.8861
         Neg Pred Value : 0.7370
             Prevalence : 0.7590
         Detection Rate : 0.7053
   Detection Prevalence : 0.7960
      Balanced Accuracy : 0.7766

       'Positive' Class : <=50K
```

# Generalized Logistic Model with Pre-processing

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

```
> confusionMatrix(predict_lm2, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9008 1158
     >50K    685 1920

               Accuracy : 0.8557
                 95% CI : (0.8495, 0.8617)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5837
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9293
            Specificity : 0.6238
         Pos Pred Value : 0.8861
         Neg Pred Value : 0.7370
             Prevalence : 0.7590
         Detection Rate : 0.7053
   Detection Prevalence : 0.7960
      Balanced Accuracy : 0.7766

       'Positive' Class : <=50K
```

# Generalized Logistic Model with Cross-validation

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

```
> confusionMatrix(predict_lm2, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9008 1158
     >50K    685 1920

               Accuracy : 0.8557
                 95% CI : (0.8495, 0.8617)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5837
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9293
            Specificity : 0.6238
         Pos Pred Value : 0.8861
         Neg Pred Value : 0.7370
             Prevalence : 0.7590
         Detection Rate : 0.7053
   Detection Prevalence : 0.7960
      Balanced Accuracy : 0.7766

       'Positive' Class : <=50K
```

# Generalized Logistic model with regularization penalty

- Parameters setting:
  - Method: glmnet
  - Pre-processing: centered and scaled
  - Cross-validation: 10-fold
  - Bootstrapped with 25 repetitions
  - Alpha range(0.1,1,0.1)
  - Lambda range(0,001,1,0.111)
- Achieved an accuracy of 85.61%, which is the best accuracy out of all the models

```
> confusionMatrix(predict_lm2, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9008 1158
     >50K    685 1920

               Accuracy : 0.8557
                 95% CI : (0.8495, 0.8617)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5837
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9293
            Specificity : 0.6238
         Pos Pred Value : 0.8861
         Neg Pred Value : 0.7370
             Prevalence : 0.7590
         Detection Rate : 0.7053
   Detection Prevalence : 0.7960
      Balanced Accuracy : 0.7766

       'Positive' Class : <=50K
```

# BEST MODEL

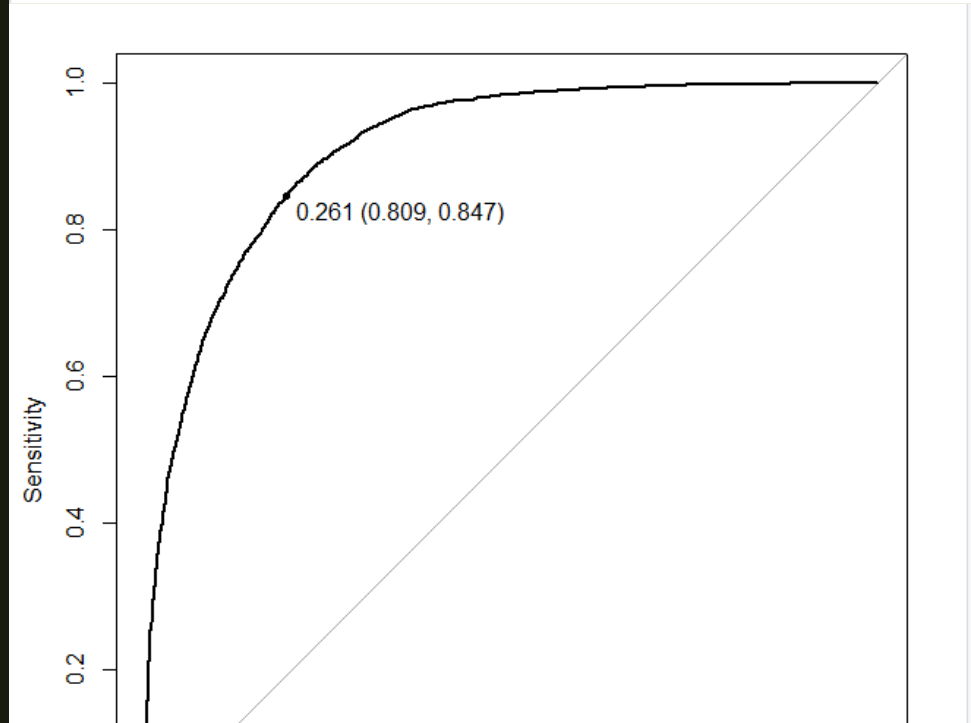
Generalized Logistic model with regularization penalty

0.5 0.001 0.8544654 0.573640974

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were alpha = 0.5 and lambda = 0.001.

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9042 1187
     >50K    651 1891

               Accuracy : 0.8561
                 95% CI : (0.8499, 0.8621)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5818
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9328
            Specificity : 0.6144
         Pos Pred Value : 0.8840
         Neg Pred Value : 0.7439
             Prevalence : 0.7590
         Detection Rate : 0.7080
   Detection Prevalence : 0.8010
      Balanced Accuracy : 0.7736
```



# IMPORTANT FEATURES

Only 20 most important variables shown (out of 48)

| | Overall |
|---|---|
| adult_data.marital.statusMarried-civ-spouse | 100.00 |
| capital.gainhigh | 94.64 |
| hours.per.weekhigh | 75.78 |
| adult_data.educationBachelors | 58.99 |
| adult_data.relationshipOwn-child | 51.09 |
| hours.per.weeknormal | 50.34 |
| adult_data.sexMale | 49.16 |
| adult_data.occupationExec-managerial | 48.47 |
| adult_data.educationMasters | 47.38 |
| adult_data.marital.statusNever-married | 46.76 |
| adult_data.educationProf-school | 38.20 |
| adult_data.educationDoctorate | 34.50 |
| adult_data.occupationProf-specialty | 32.55 |
| adult_data.relationshipWife | 30.46 |
| adult_data.educationSome-college | 28.61 |
| capital.lossnormal | 28.44 |
| adult_data.occupationSales | 24.82 |
| adult_data.occupationOther-service | 24.28 |

```
coef[order(coef[, 4]), ]
```

| | Estimate | Std. Error | z value | Pr(>\|z\| |
|---|---|---|---|---|
| pital.gainhigh | 0.76083461 | 0.03080382 | 24.699358 | 1.086611e-13 |
| pital.lossnormal | 0.21887492 | 0.01967516 | 11.124430 | 9.541372e-2 |
| dult_data.educationProf-school` | 0.35165555 | 0.03406509 | 10.323048 | 5.543462e-2 |
| ult_data.relationshipWife | 0.28974542 | 0.02825374 | 10.255118 | 1.122389e-2 |
| ult_data.educationDoctorate | 0.30954444 | 0.03034081 | 10.202247 | 1.937373e-2 |
| urs.per.weekhigh | 0.59741772 | 0.06232079 | 9.586170 | 9.141578e-2 |
| ult_data.educationMasters | 0.44407912 | 0.04704252 | 9.439953 | 3.729481e-2 |
| dult_data.occupationExec-managerial` | 0.32586657 | 0.03767299 | 8.649873 | 5.155693e-1 |
| ult_data.sexMale | 0.40605500 | 0.04769121 | 8.514253 | 1.676676e-1 |
| ult_data.educationBachelors | 0.60866554 | 0.07402568 | 8.222357 | 1.995427e-1 |
| dult_data.marital.statusMarried-civ-spouse` | 1.30044460 | 0.18150962 | 7.164604 | 7.801148e-1 |
| urs.per.weeknormal | 0.38598060 | 0.06513407 | 5.925940 | 3.105155e-0 |
| dult_data.occupationProf-specialty` | 0.22195275 | 0.04060627 | 5.465972 | 4.603774e-0 |
| dult_data.educationAssoc-voc` | 0.22245764 | 0.04351188 | 5.112572 | 3.178017e-0 |
| dult_data.occupationTech-support` | 0.12590200 | 0.02633686 | 4.780449 | 1.749040e-0 |
| dult_data.educationSome-college` | 0.38357436 | 0.08253152 | 4.647610 | 3.358024e-0 |
| dult_data.educationAssoc-acdm` | 0.18727487 | 0.04072257 | 4.598797 | 4.249368e-0 |
| ult_data.occupationSales | 0.16504761 | 0.03737187 | 4.416360 | 1.003769e-0 |
| dult_data.occupationFarming-fishing` | -0.13131800 | 0.03138261 | -4.184419 | 2.858958e-0 |
| dult_data.occupationOther-service` | -0.20694536 | 0.05035730 | -4.109541 | 3.964471e-0 |
| dult_data.occupationProtective-serv` | 0.09798208 | 0.02429490 | 4.033031 | 5.506212e-0 |
| dult_data.marital.statusMarried-AF-spouse` | 0.07865652 | 0.02019242 | 3.895348 | 9.805769e-0 |
| urs.per.weeksuper | 0.08342491 | 0.02283648 | 3.653143 | 2.590500e-0 |
| dult_data.workclassFederal-gov` | 0.12322112 | 0.03461770 | 3.559483 | 3.715858e-0 |
| dult_data.educationHS-grad` | 0.30028592 | 0.09103149 | 3.298704 | 9.713231e-0 |
| ult_data.marital.statusWidowed | 0.09841204 | 0.03343011 | 2.943814 | 3.241947e-0 |
| dult_data.marital.statusNever-married` | -0.14948449 | 0.05219694 | -2.863856 | 4.185187e-0 |
| dult_data.occupationHandlers-cleaners` | -0.10484833 | 0.03852661 | -2.721452 | 6.499576e-0 |
| dult_data.relationshipNot-in-family` | 0.42307920 | 0.15816510 | 2.674921 | 7.474677e-0 |
| dult_data.workclassSelf-emp-inc` | 0.07456133 | 0.03585144 | 2.079730 | 3.755028e-0 |

Based on Association Rules and output from the previous model

We selected Key parameters below:

Sex, Native.country, Martial.status, Capital, Relationship

Logistic Regression

10-fold cross validation

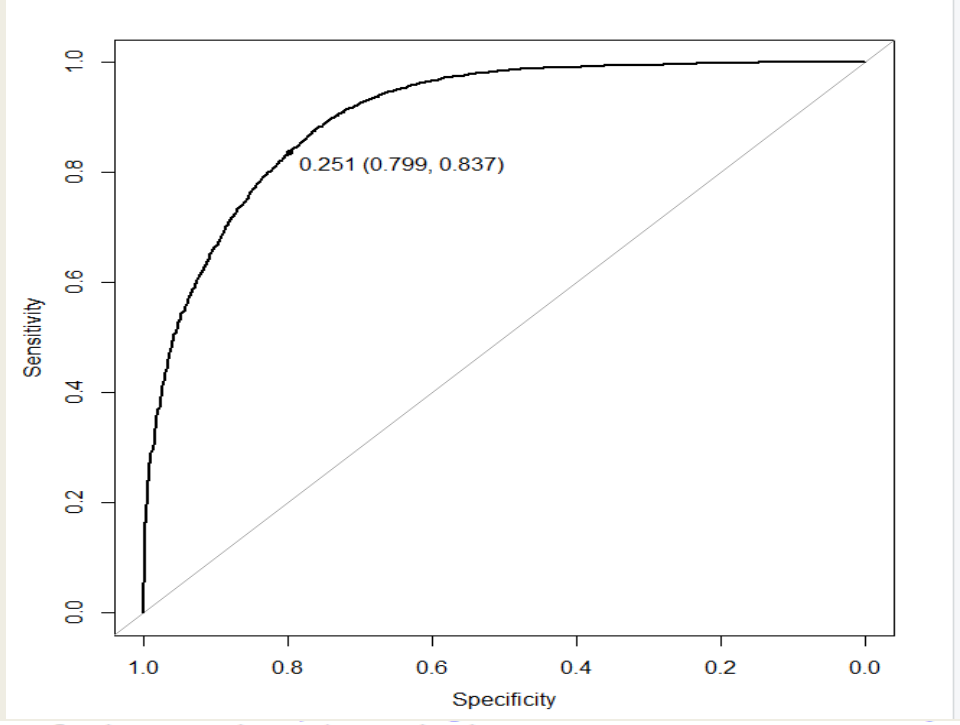
Regularization Penalty (Alpha (0.1,1,0.1) lambda(0.001,1,0.111)

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

Accuracy was used to select the optimal model using the l argest value.

The final values used for the model were alpha = 0.5 and l ambda = 0.001.




```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9036 1273
     >50K    657 1805

               Accuracy : 0.8489
                 95% CI : (0.8425, 0.855)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5567
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9322
            Specificity : 0.5864
         Pos Pred Value : 0.8765
         Neg Pred Value : 0.7331
             Prevalence : 0.7590
         Detection Rate : 0.7075
   Detection Prevalence : 0.8072
      Balanced Accuracy : 0.7593
```

Logistic Regression

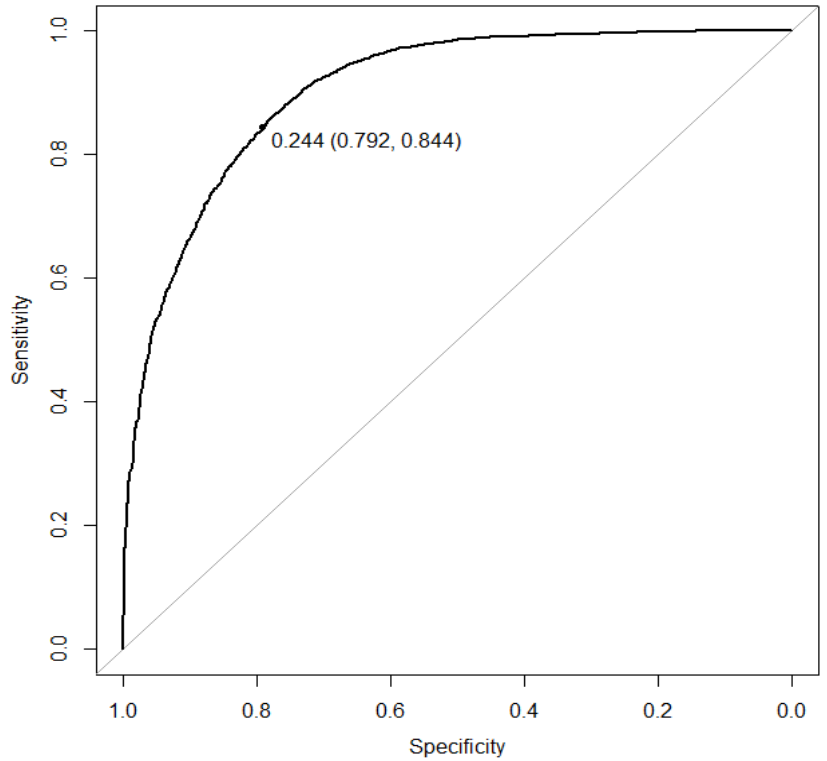
10-fold cross validation

Regularization Penalty (Alpha (0.1,1,0.1)
lambda(0.001,1,0.111)

Selected input
adult_data.salary ~ capital.gain + capital.loss
+adult_data.education
+adult_data.relationship+hours.per.week+adult_data.occ
upation+adult_data.sex+adult_data.marital.status

1.0 0.001 0.8499761 0.55235158

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were alpha = 1 and lambda = 0.001.




```
> confusionMatrix(predict_lg6, testdata$adult_data.salary)
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  9045 1284
     >50K    648 1794

               Accuracy : 0.8487
                 95% CI : (0.8424, 0.8549)
    No Information Rate : 0.759
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5551
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9331
            Specificity : 0.5828
         Pos Pred Value : 0.8757
         Neg Pred Value : 0.7346
             Prevalence : 0.7590
         Detection Rate : 0.7082
   Detection Prevalence : 0.8088
      Balanced Accuracy : 0.7580

       'Positive' Class : <=50K
```

# BUSINESS RECOMMENDATIONS

# Business problem

| Human Resource | Context | Issues |
| --- | --- | --- |
| • Salary mapping in accordance with employee portfolio<br>• Dynamic higher-management compensation adjustment<br>• Record management for socio-economic attributes and performance KPIs | • Retention/attrition rate<br>• Compensation allocation<br>• Past/present employee history<br>• ERP integration<br>• CRM Integration<br>• Predictive capabilities in the Human Resource solution | • Relational databases not inter-connected<br>• Missing data-points<br>• Incorrect data-points considered in the system, leading to wrong predictions<br>• In-efficient resource allocation<br>• Customer journey with Sales/Marketing team not mapped and modeled |

# Challenges deep-dive

Challenge 1 → Challenge 2 → Challenge 3

**Garbage data**

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

**Lack of Technological/Managerial Capability**

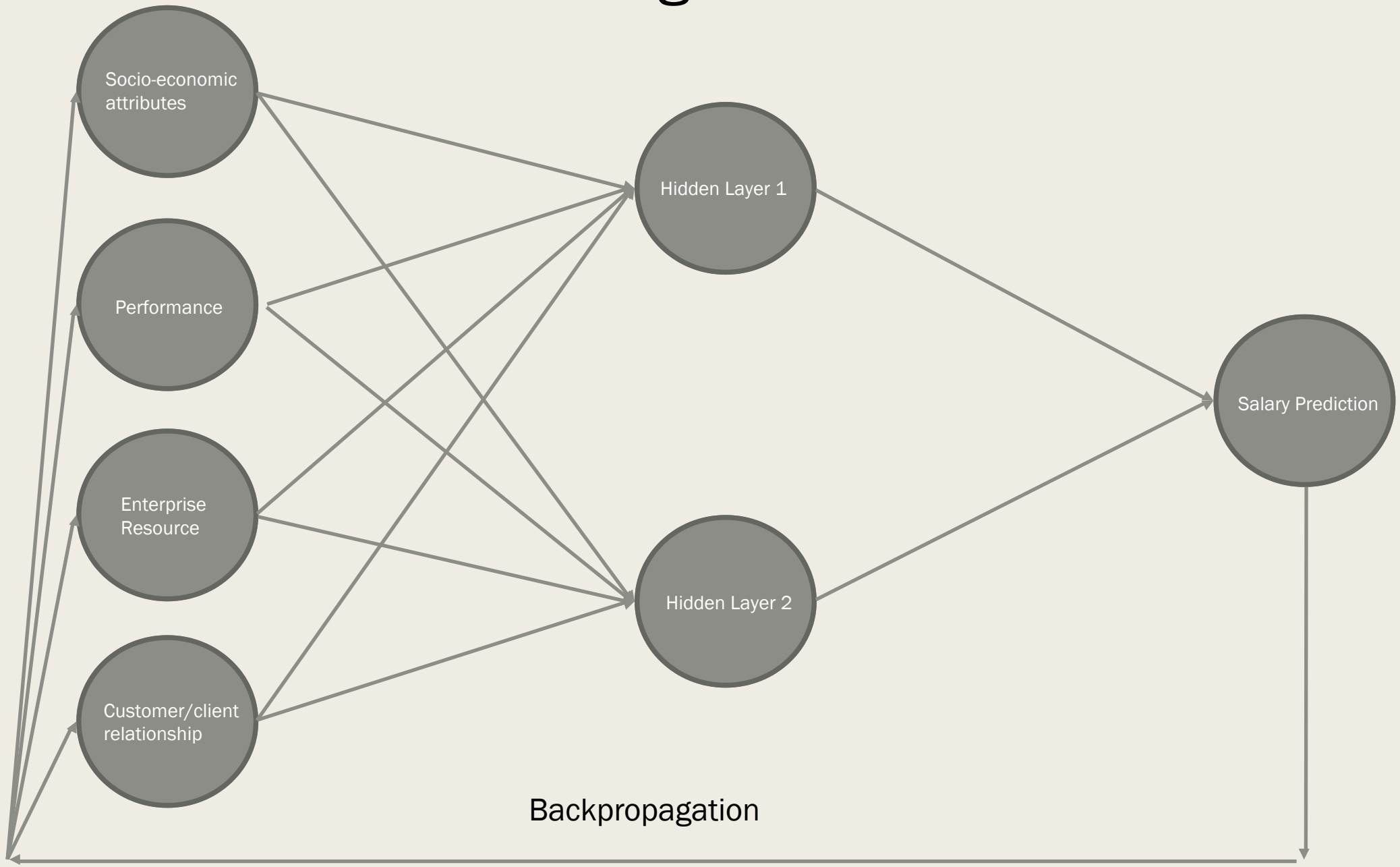
- Lack of cloud-infrastructure for big-data analytics
- Lack of design and reasoning skills when creating survey
- Inability to engage employees in filling out surveys
- Internal politics in the organization leads to generation of wrong or sparse data points, due to favouritism
- Lack of knowledge in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

**Unclear Plan of Action**

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

# IMPLEMENTATION

# Architecture of the Intelligent HR solution



# Technical Capability Roadmap

## 1-year goal

Data Procurement and engineering:

- Collecting data
- Data cleaning and filtering
- Noise reduction
- Correlation Tests
- Heteroscedasticity
- PCA/clustering
- Statistical significance tests

Data Storage:

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

## 2-year goal

Traditional Machine learning:

- Linear/Logistic regression
- SVM
- KNN
- Naive Bayes
- Decision Trees

Sampling and Ensemble Learning:

- Sampling tests
- Bootstrap aggregating
- Boosting
- Bucket of Models
- Stacking
- ROC curve and Kappa metric evaluation

## 4-year goal

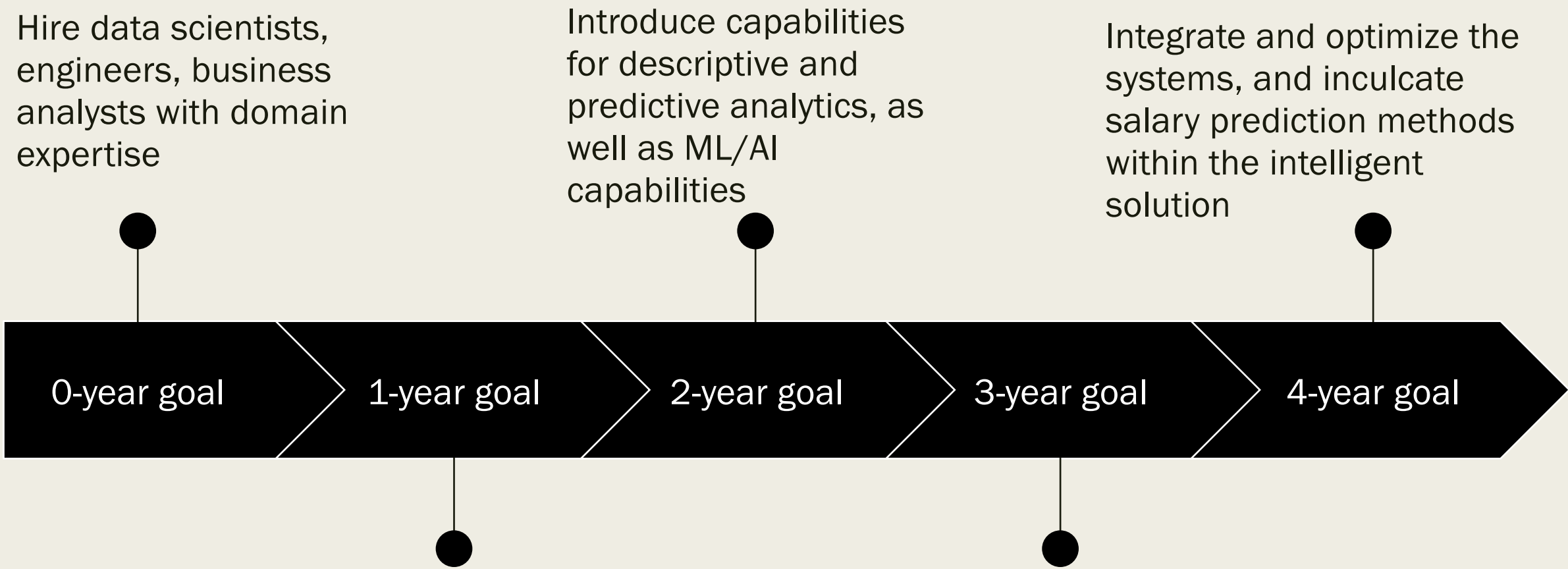
Knowledge infrastructure and pricing strategy:

- Knowledge centre development
- Content strategy for thought leadership
- CRM integration strategy
- ERP integration strategy
- Salary-allocation strategy

Artificial intelligence capabilities:

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

# Strategic Roadmap



# THANK YOU!
# Any Questions?