
R


python


sas

The Comparison of Income Level Prediction Models

Qiudong Deng

Project Goal

This project aims to select the best model from logistic regression, random forest , support vector machine, decision tree and gradient boosting models to predict if the income level of an individual in the US is greater than or less than $50,000 based on the information available for the individual from the census data.

Area under the curve (AUC) of receiver operating characteristic (ROC) and confusion matrix as selection criterions

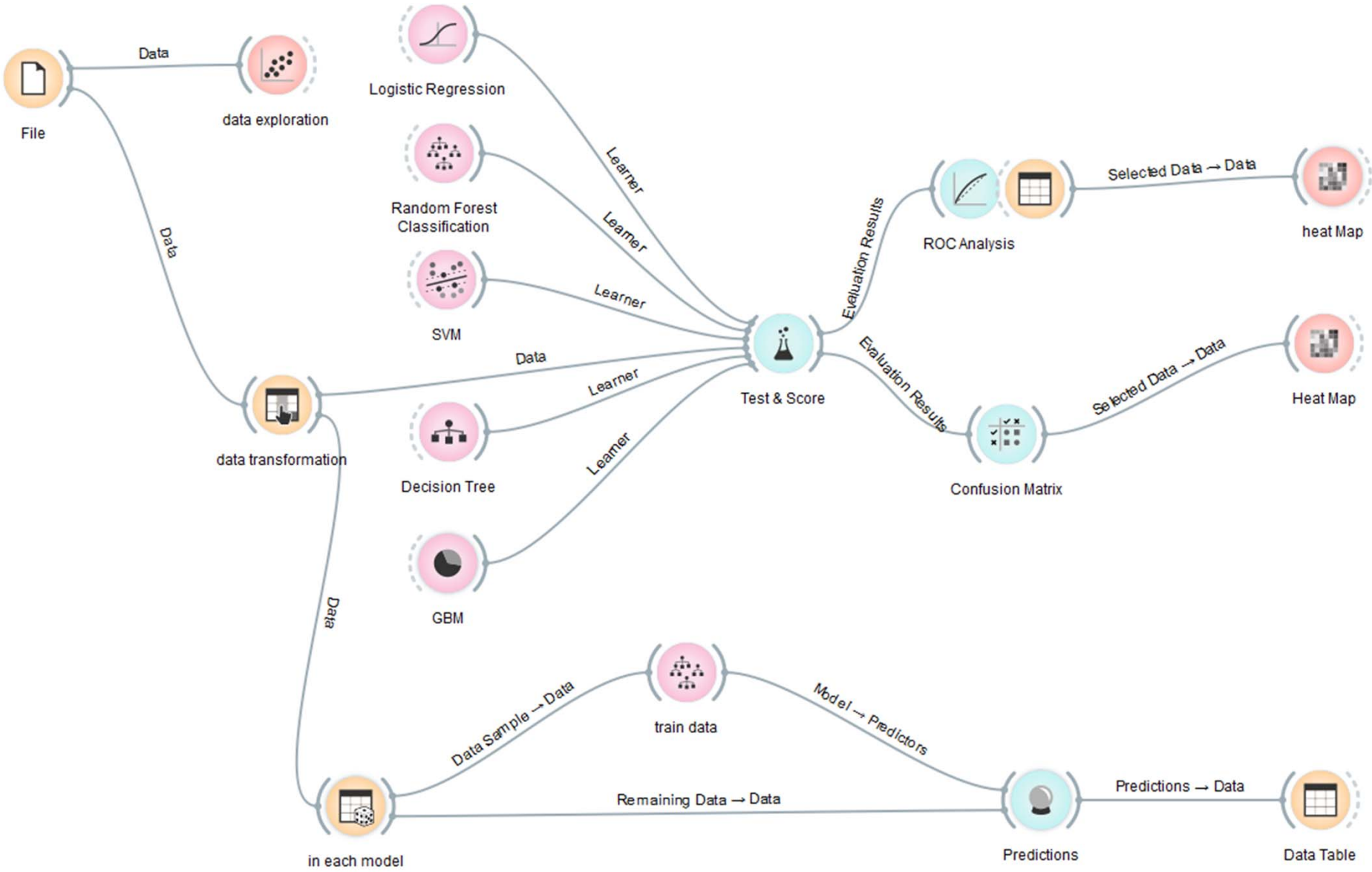
The dependent (target) variable for this project is binary (income level>50K or <=50K).

Data

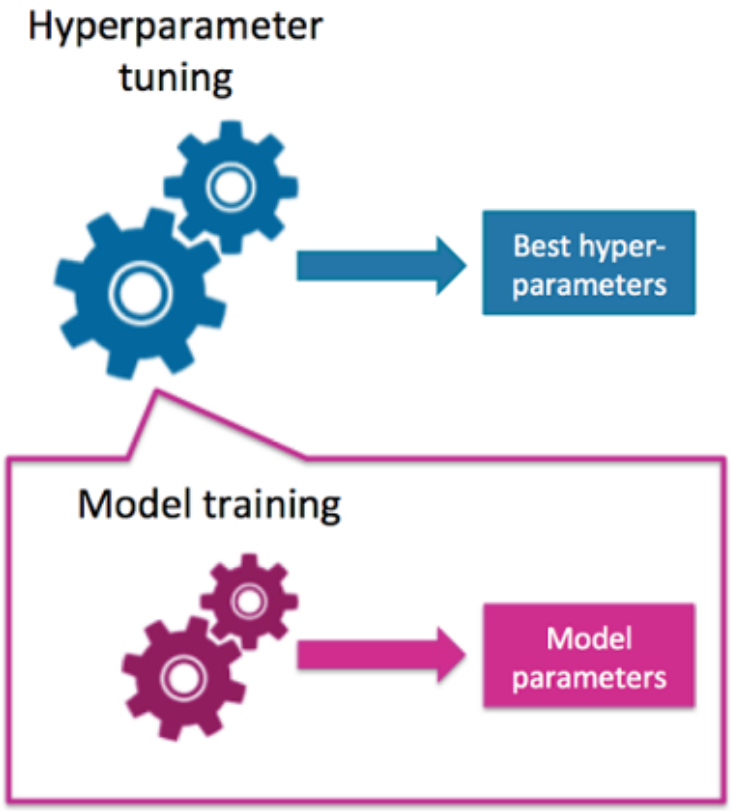
- This dataset named “Census Income Data Set” can be downloaded from the public website (http://archive.ics.uci.edu/ml/datasets/Census+Income).

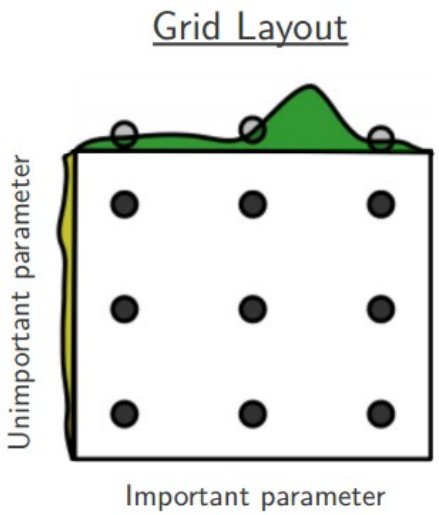
- This dataset has 48,842 instances and 14 independent variables that were extracted from the 1994 census data.

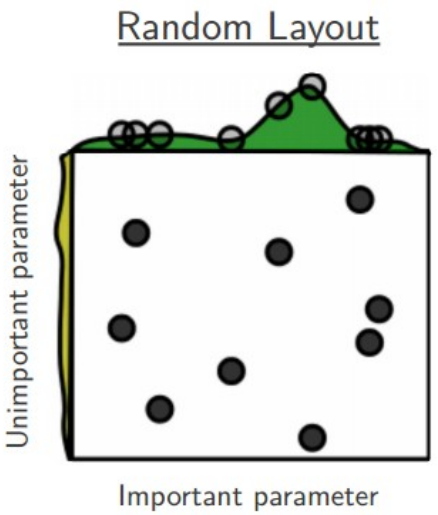
Workflow

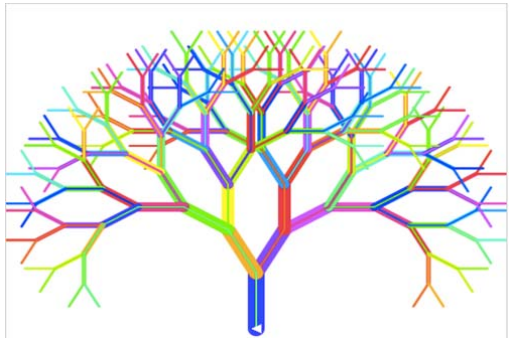

File
Data
data exploration
Data
data transformation
Logistic Regression
Random Forest Classification
SVM
Decision Tree
GBM
Learner
Learner
Learner
Data
Learner
Learner
Test & Score
Evaluation Results
ROC Analysis
Selected Data → Data
heat Map
Evaluation Results
Confusion Matrix
Selected Data → Data
Heat Map
Data
in each model
Data Sample → Data
train data
Model → Predictors
Remaining Data → Data
Predictions
Predictions → Data
Data Table

Two Layouts for Hyper-parameter Optimization in Python Scikit-learn Library



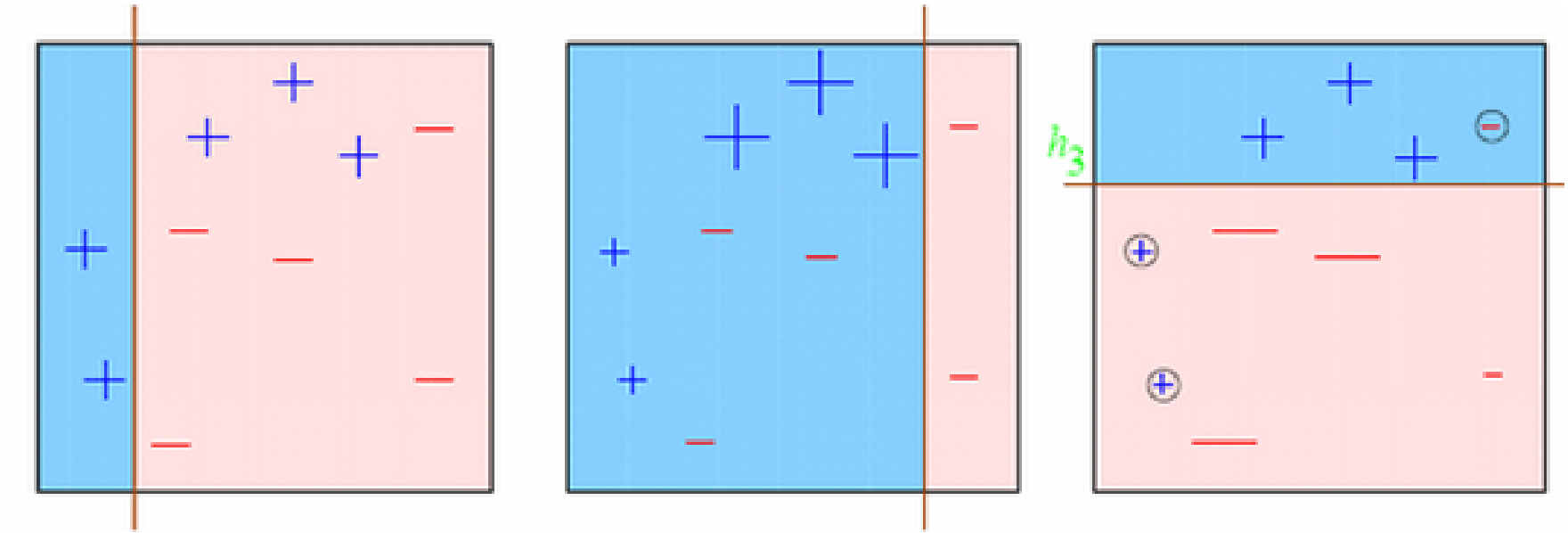






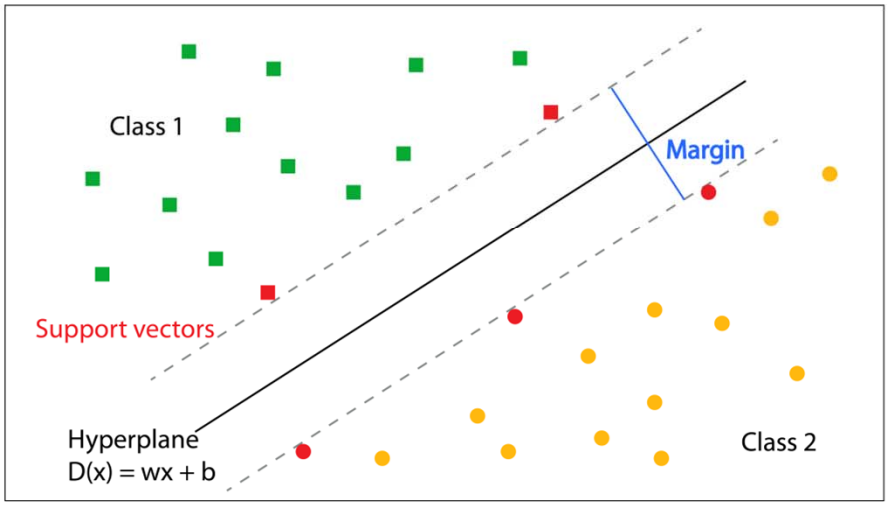
Random Forest

Gradient Boosting Machines (GBM)

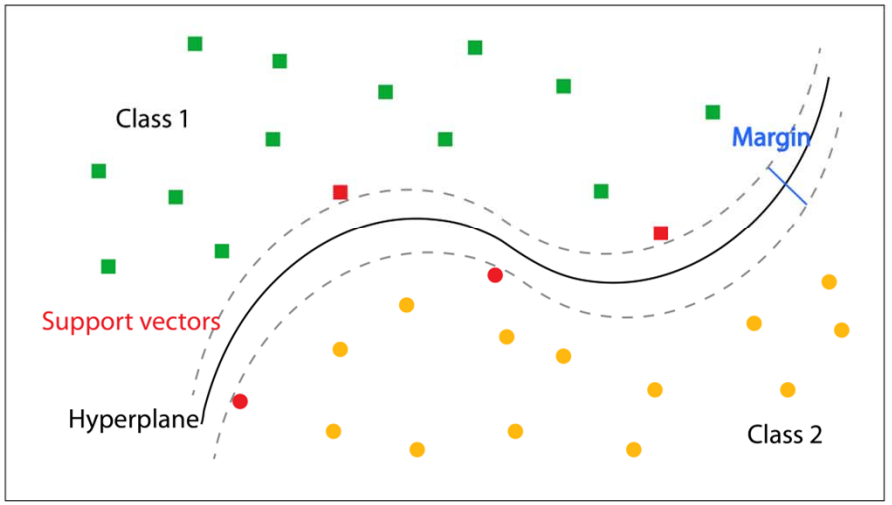


Support Vector Machines (SVM)

A. Linear separation



B. Non-linear separation



Confusion Matrix

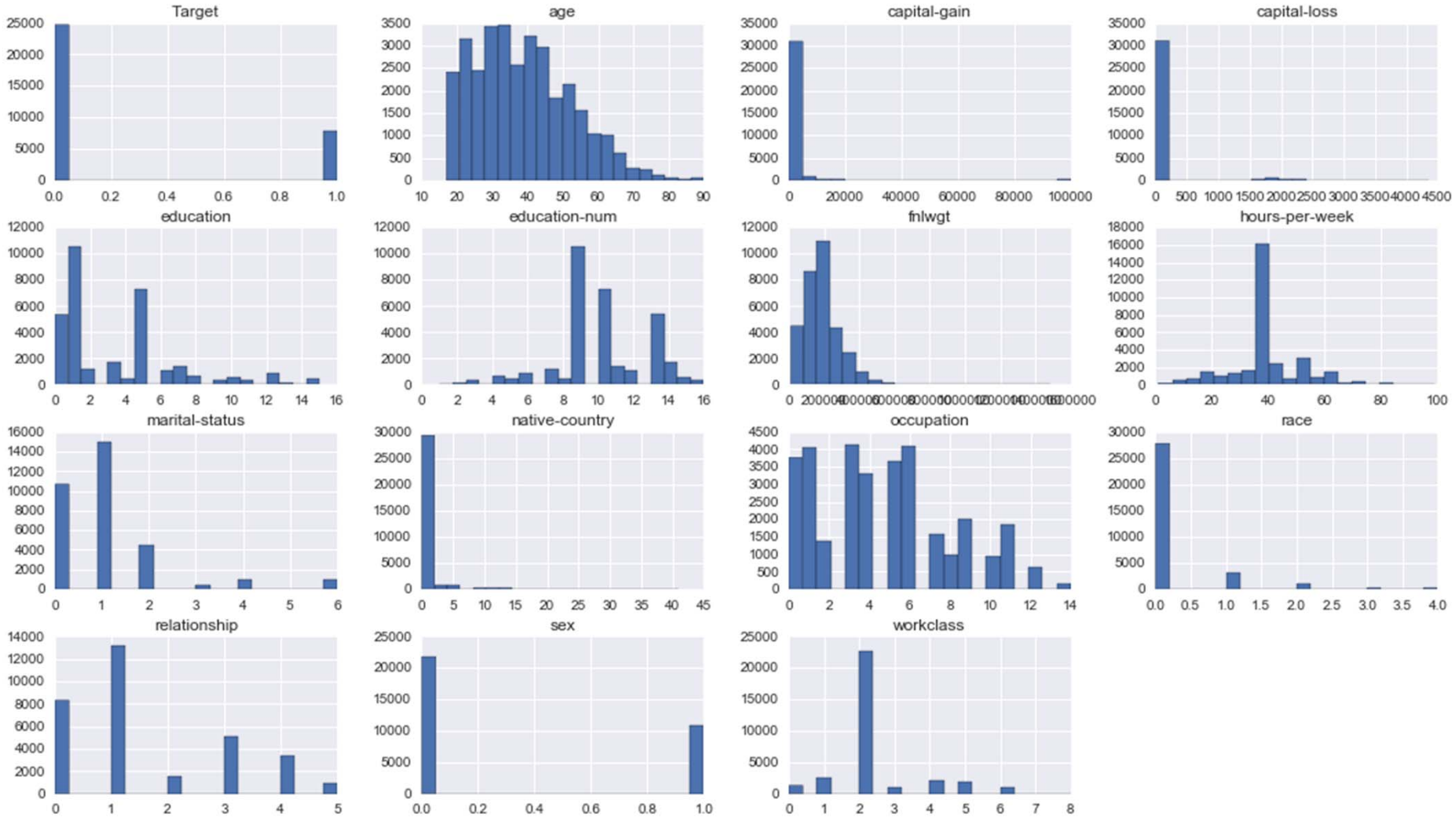
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$ =>Sensitivity

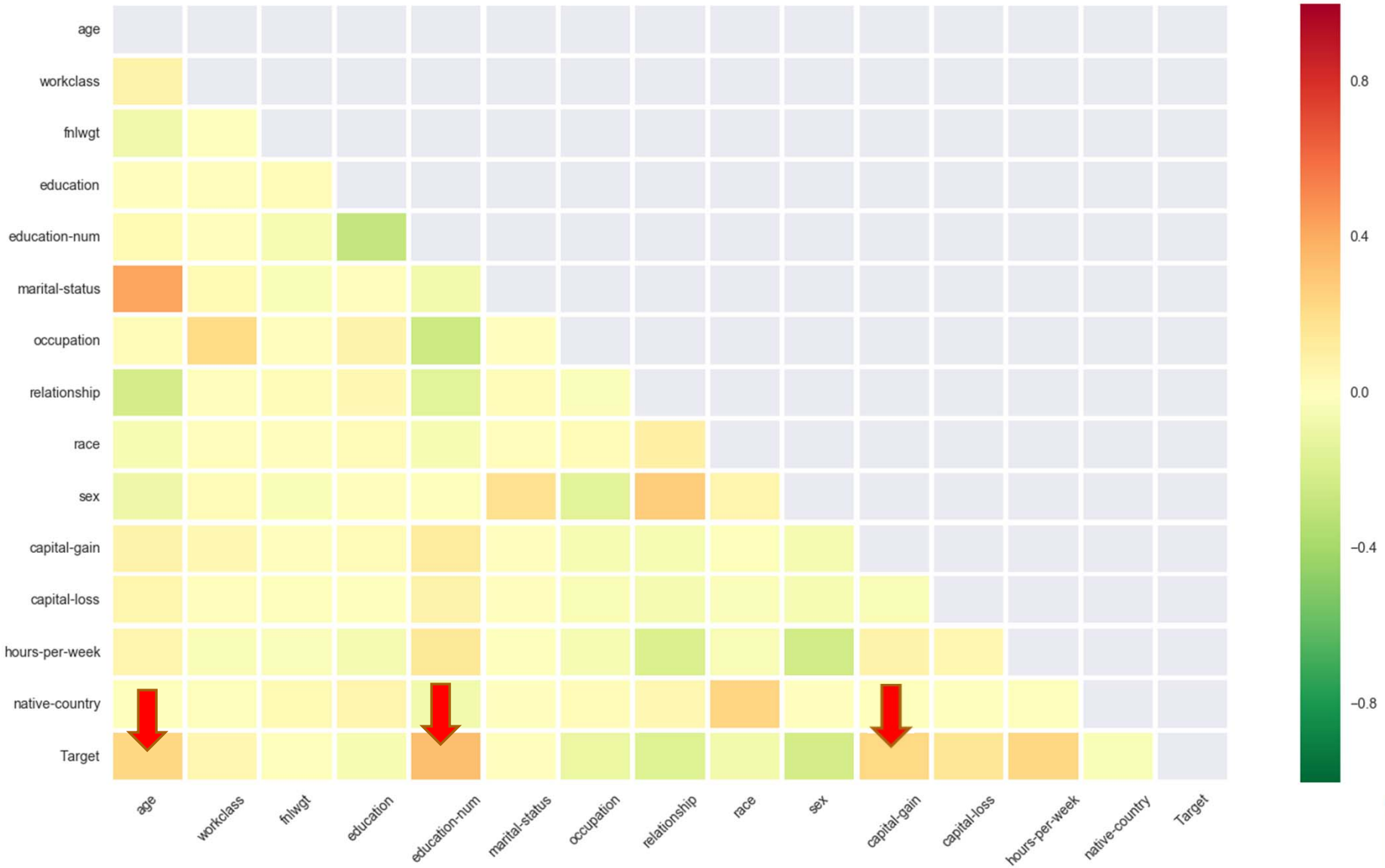
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$ => Weighted average of the precision and recall

The higher the better!

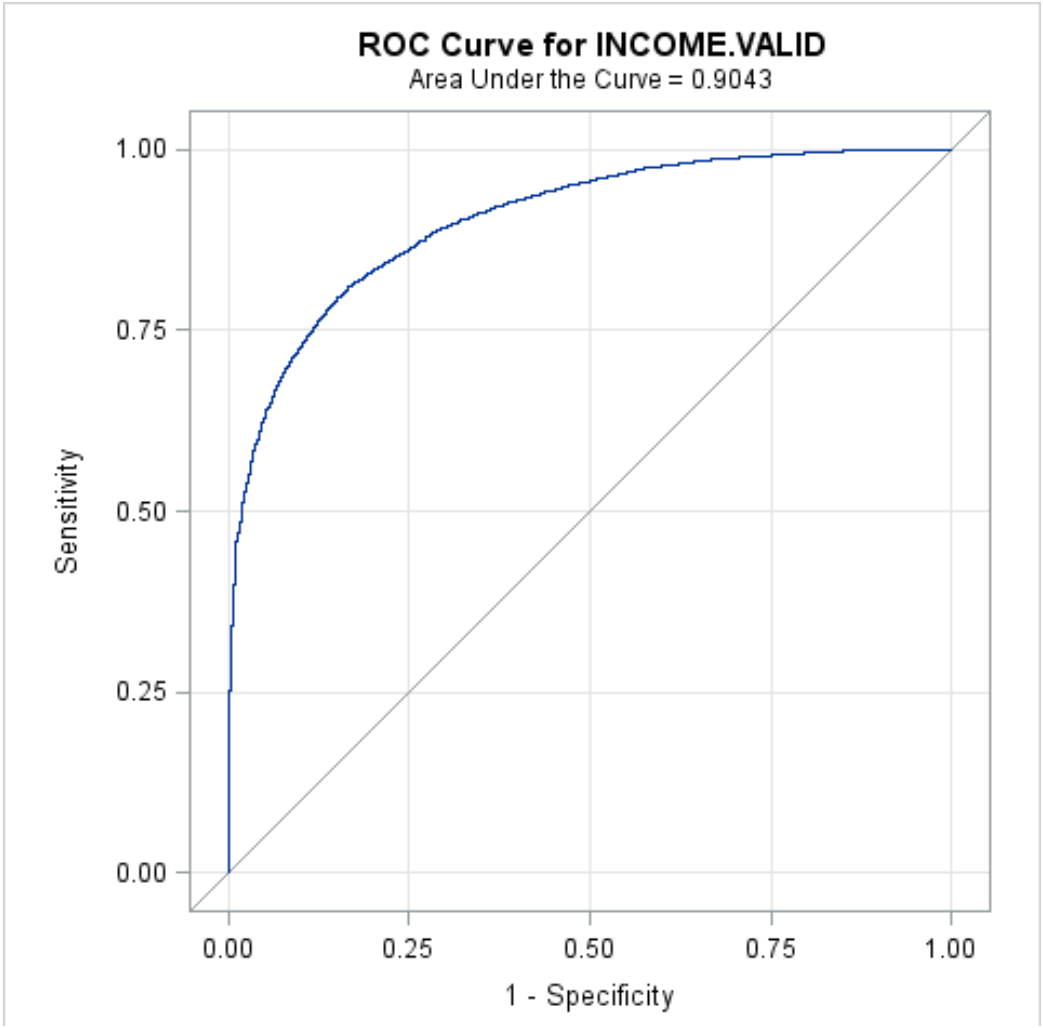
Distribution of Variables



Correlation Matrix



ROC Curve for Logistic Regression

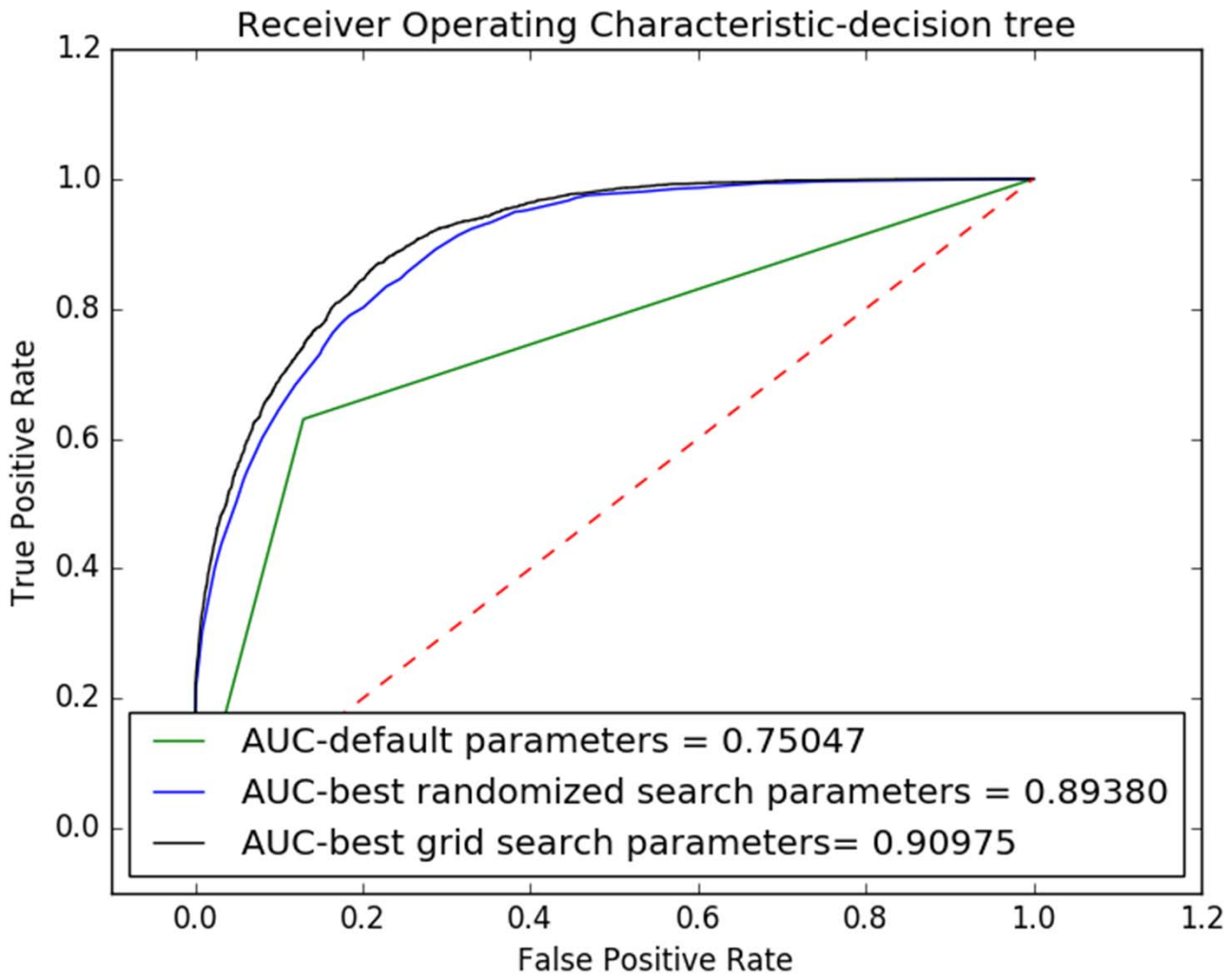


The LOGISTIC Procedure

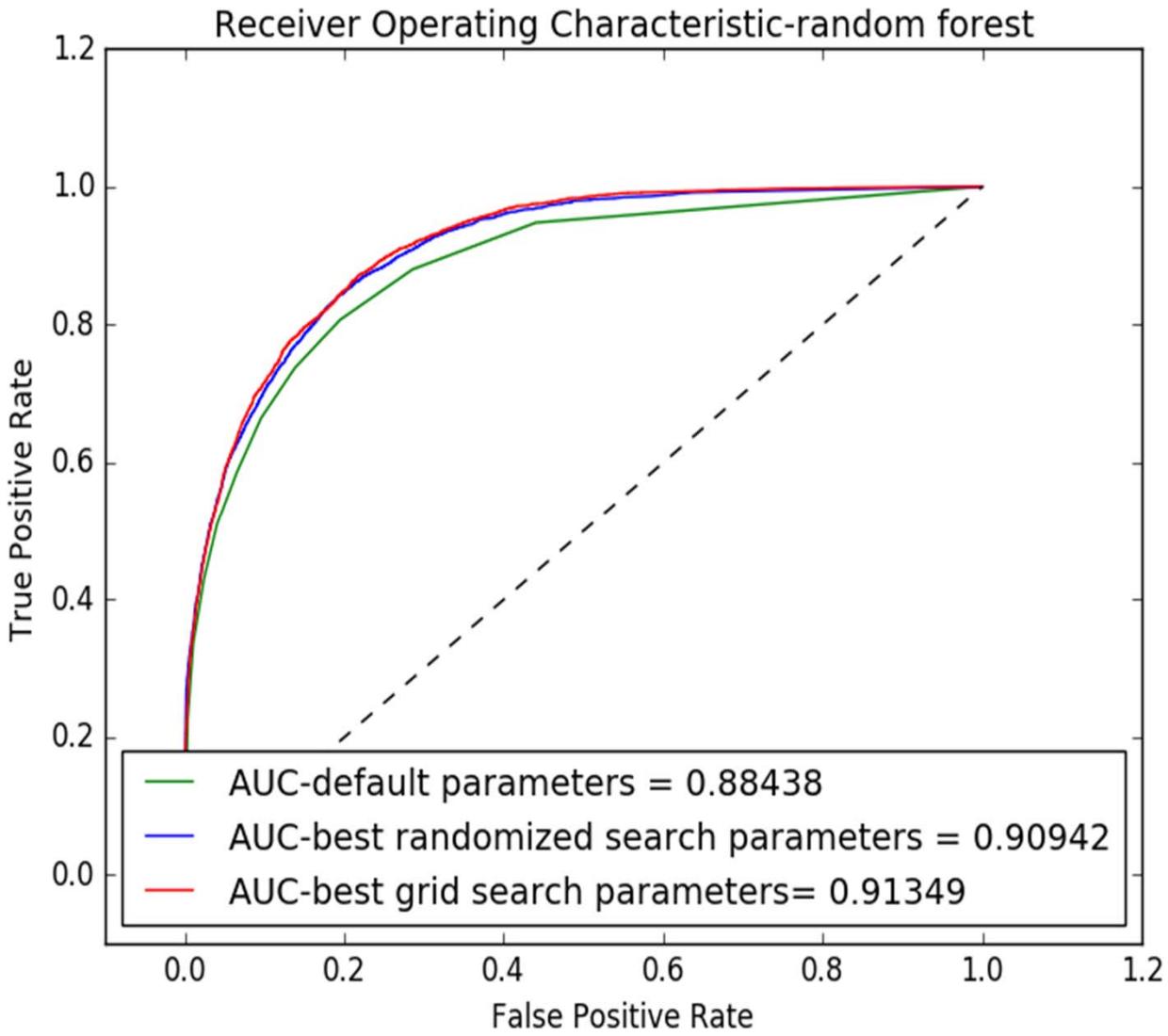
Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
INCOME.VALID	15473	-5055.0	0.1529	10222.08	10222.49	10650.3	10650.3	0.37568	0.556412	0.904259	0.104765
INCOME.TRAIN	15245	-4951.9	0.1522	10015.75	10016.17	10443.14	10443.14	0.375586	0.557324	0.90591	0.103962

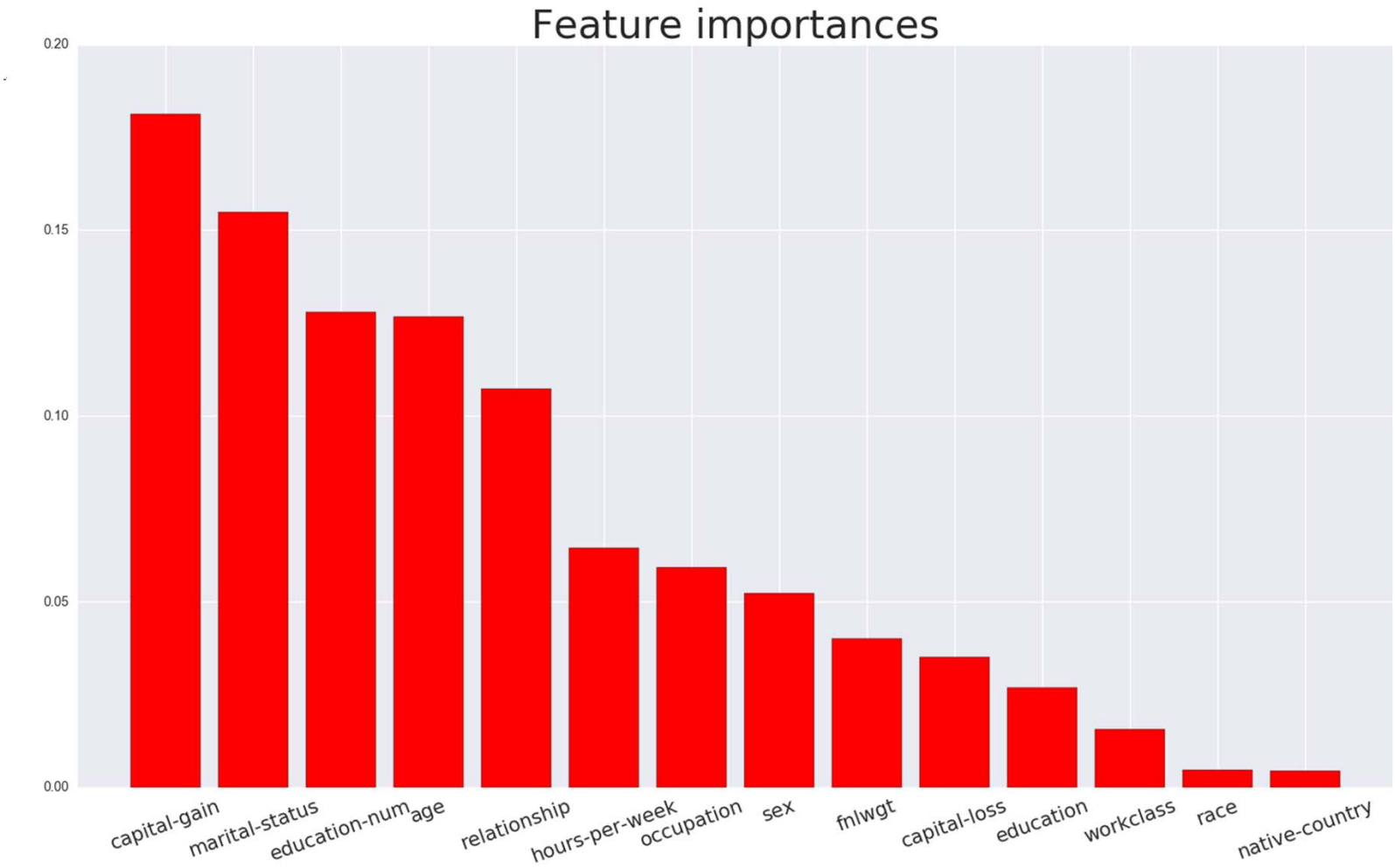
AUC Comparisons in Decision Tree



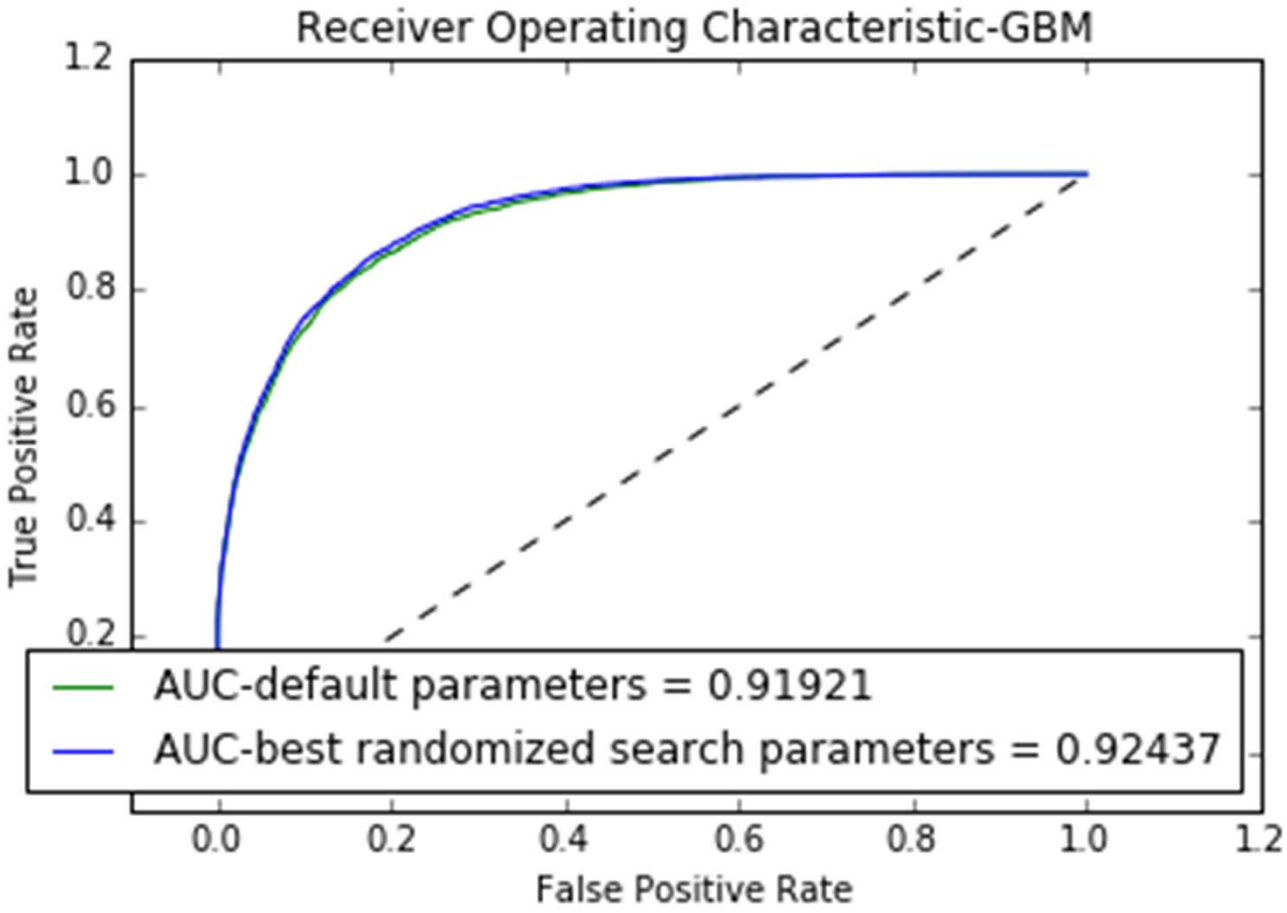
AUC Comparisons in Random Forest



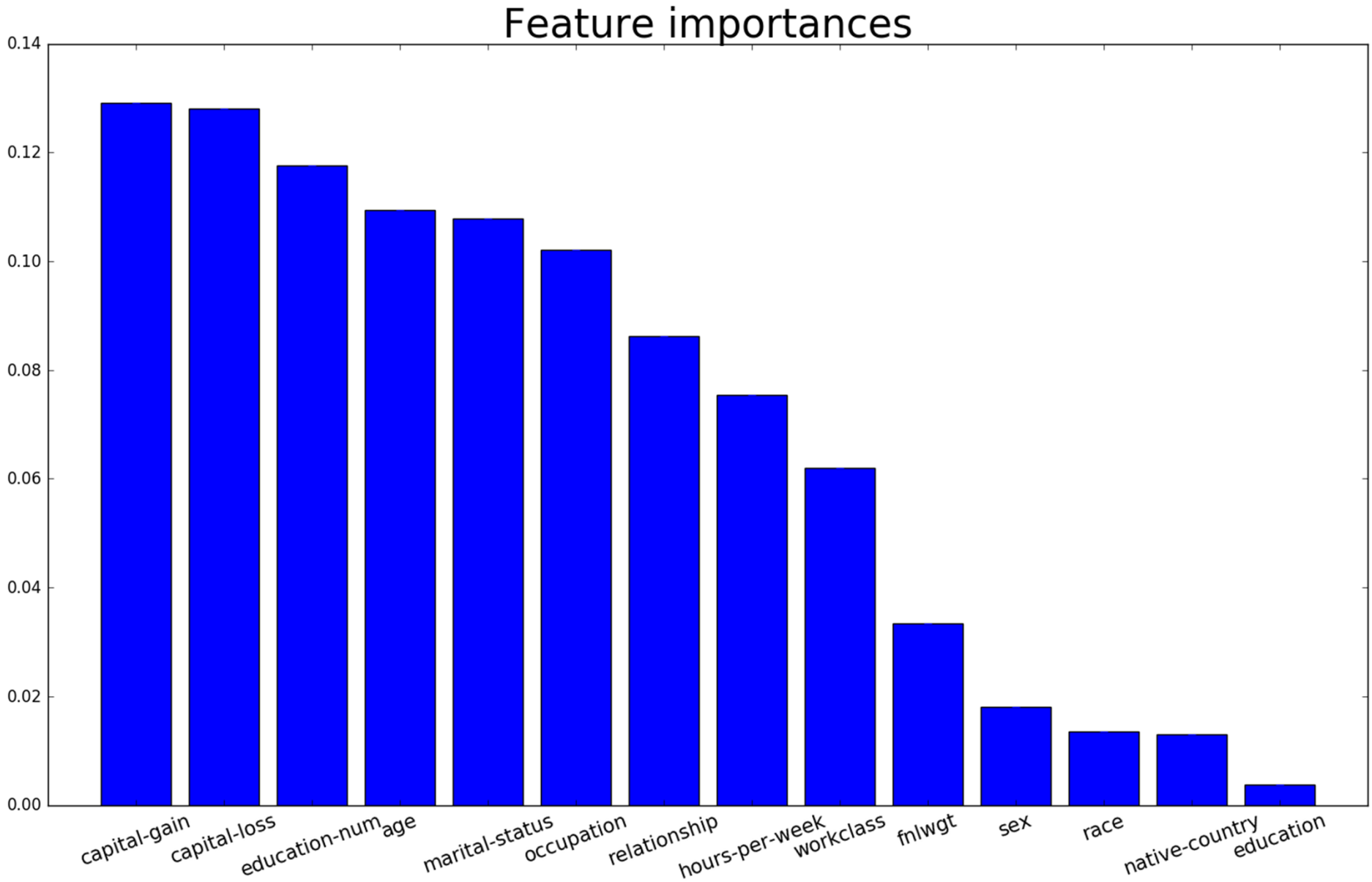
Importance Rank of Features Found in Grid Search in Random Forest



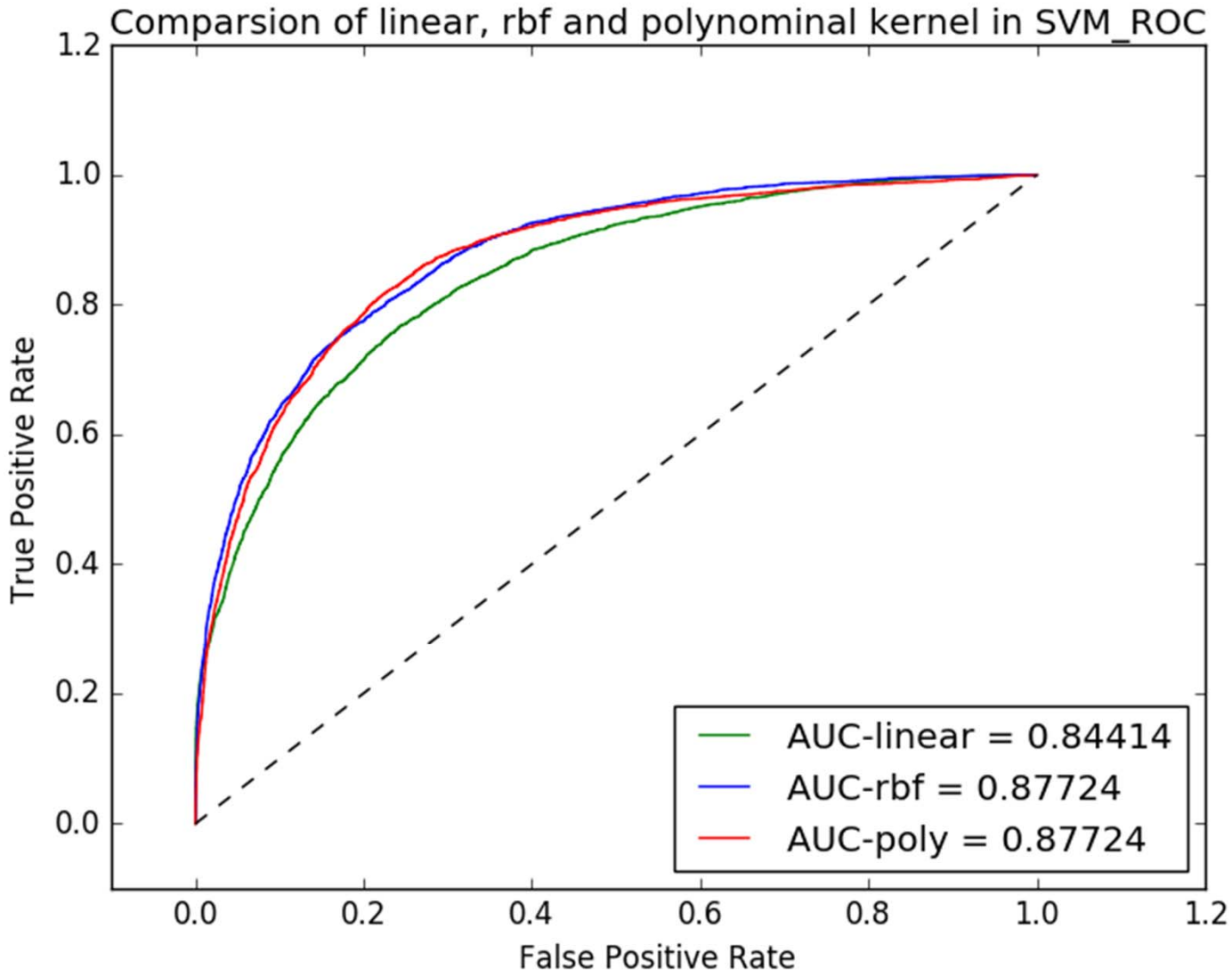
AUC Comparisons in GBM



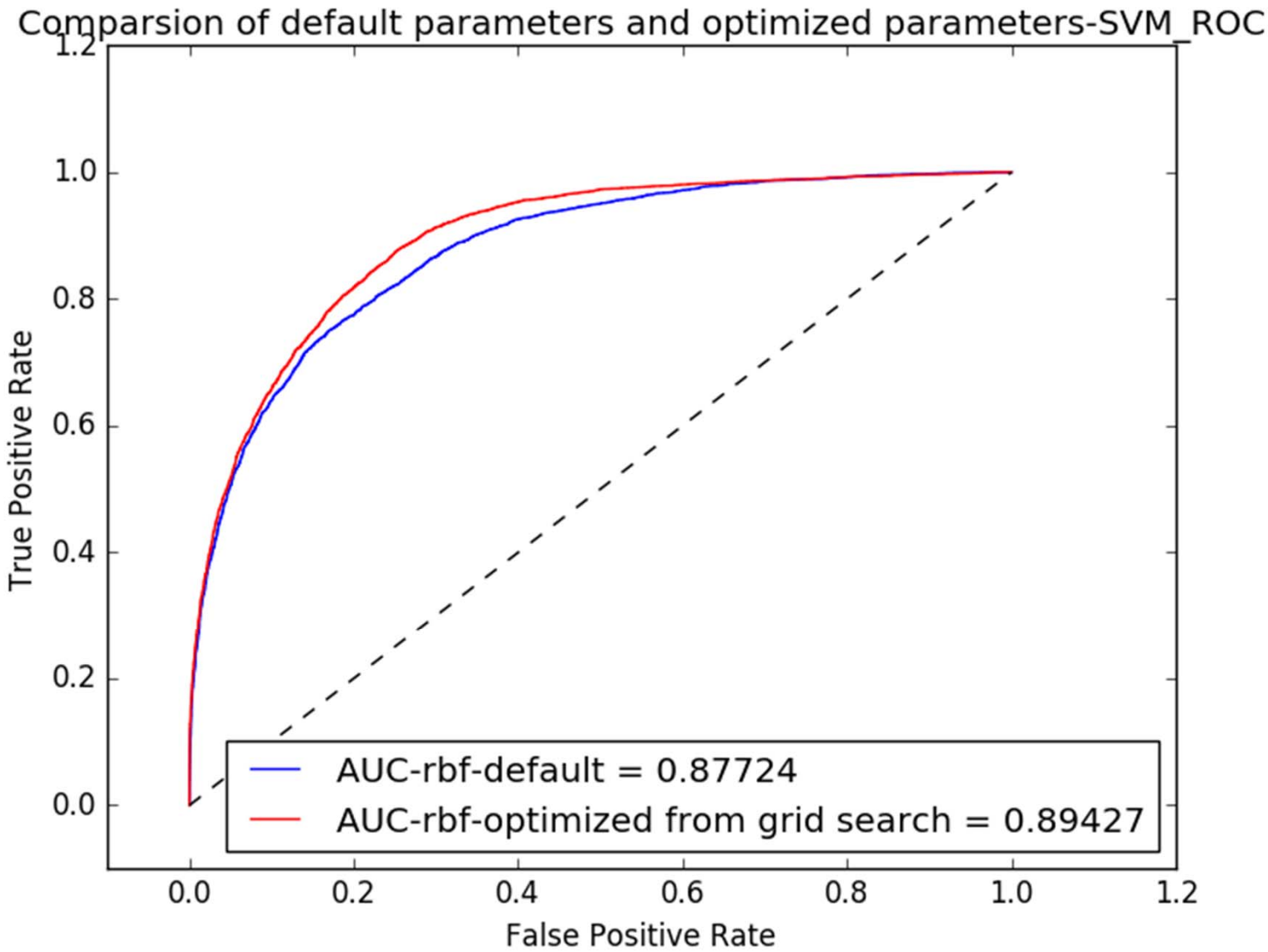
Importance Rank of Features Found in GBM



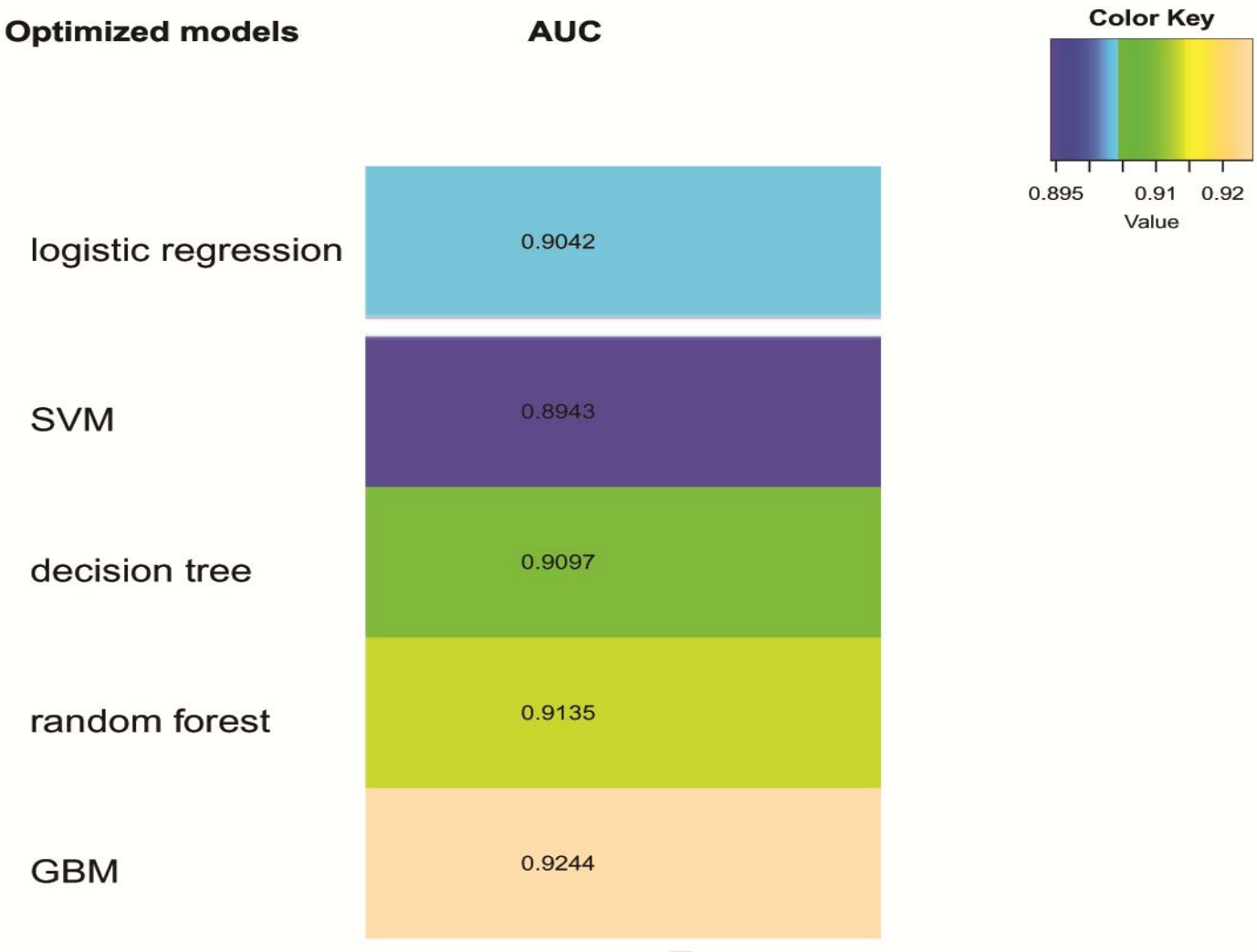
Comparison Linear and Non-linear Kernel in SVM Using Default Parameters



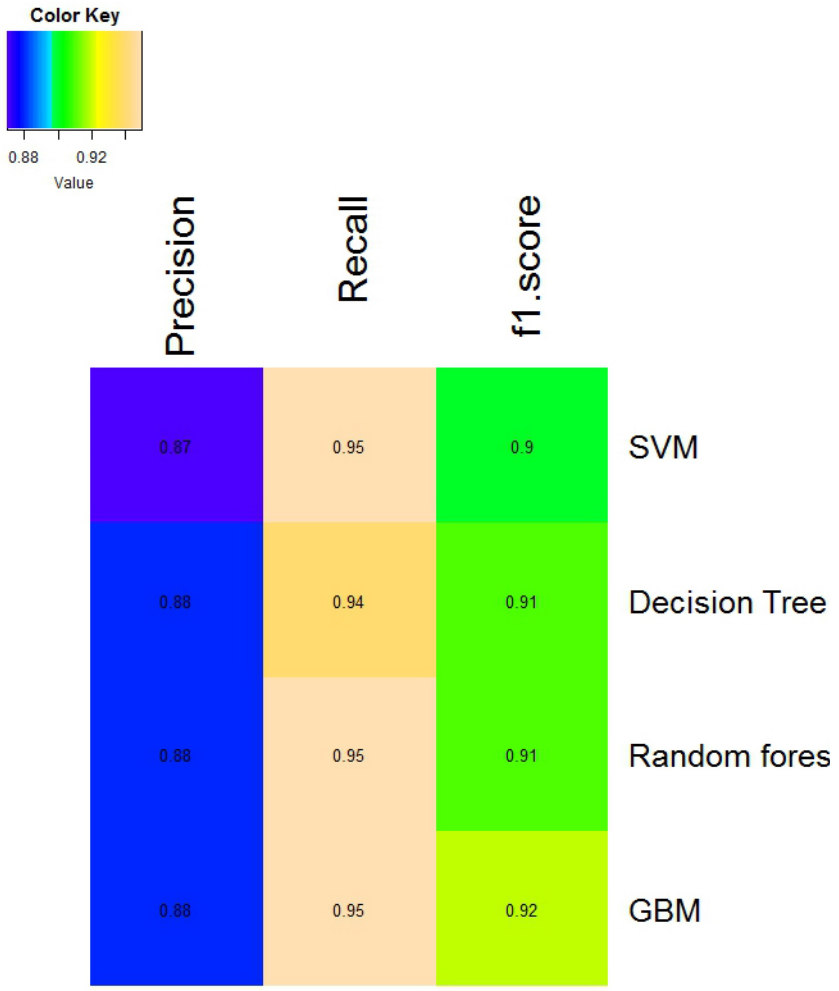
Comparison Default and Optimized Parameters -SVM

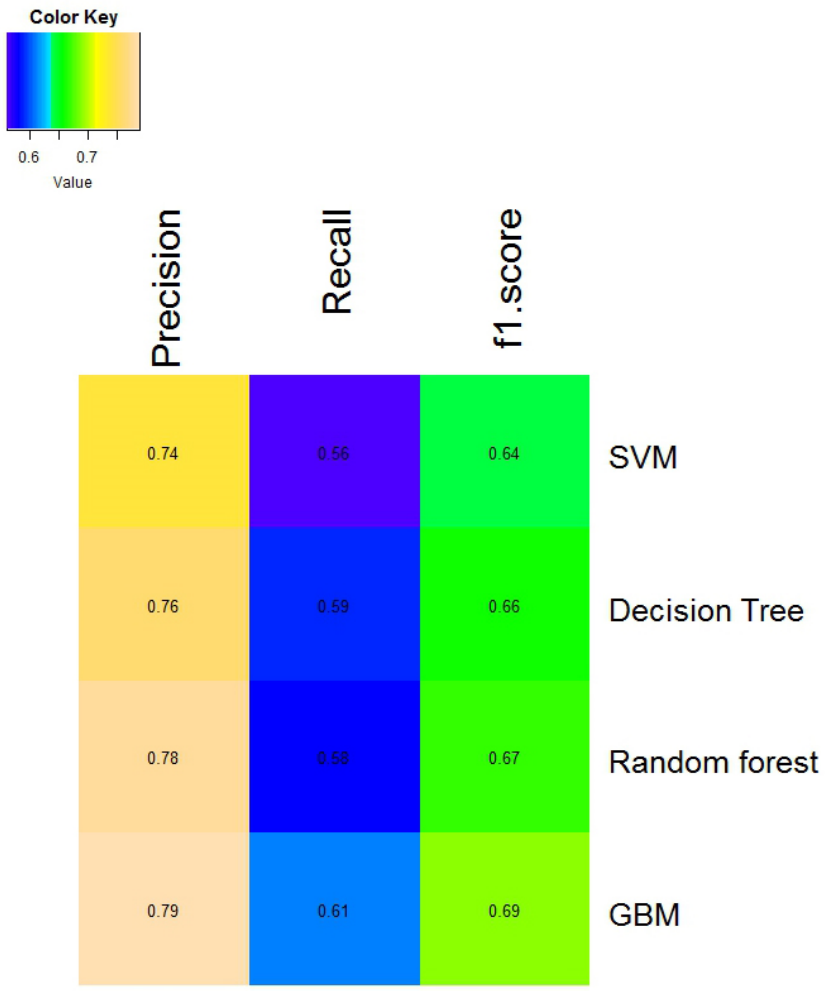


AUC Comparison of Optimized Models



Confusion Matrix Comparison on Optimized Models





Performance on predication of =<50K

>50K

Conclusions

1. AUC and confusion matrix both showed that optimized GBM model has the best performance in all models built in this project on this dataset.

2. We can do importance rank of features in GBM and random forest models. Both models showed capital gain is the most important feature to determine income level.