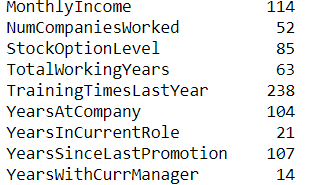
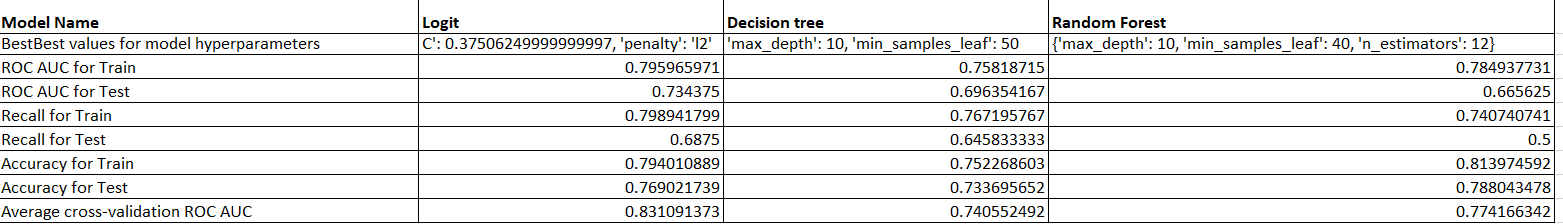
1.There were no missing values in data. Here are columns containing outliers.

2. I used winsorization, it is a transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers, it replaces outliers by max or min depending on which critical region the outlier is. In python shell you may see its increase accuracy a little bit. We had unnecessary column employeenumber so it is dropped and I drop some correlated variables in for Logistic regression.

3.

4.If I have to choose one of these models I would prefer Logit model because of following reasons.

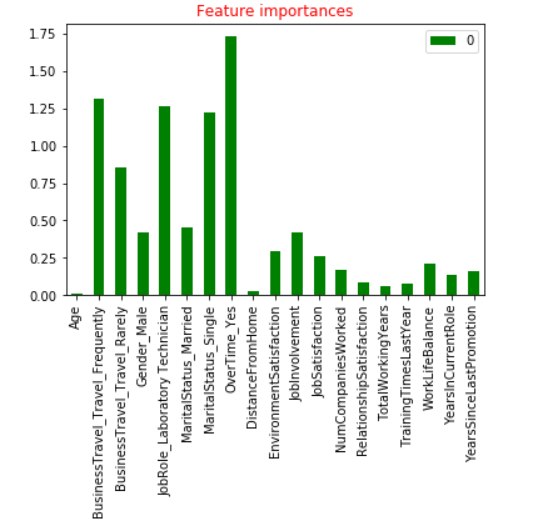
* 1.My priority is to distinguish between 0 and 1 the roc\_auc provides that score. Average score of Roc\_AUC is the biggest among these models. And not only this score other matrix scores are bigger than scores provided by Decision tree and Random forest except accuracy for train, in that case Random Forest provides bigger score but it can be due to low recall.
* 2. Model is comparably lower bias and have lower variance that the other two models.

PS. Overall the method is not good because it does not reach benchmark 0.84.

Important features are following if we state alpha=0.05



5.Note for visualization I used logit.coefs for those variables whose stats.models logits p-values were less than alpha because sklearns LogisticRegression does not provide features importance.

.

6. **Problem 2**

Raymond Reddington (known as well as Red) is a data scientist who aims to predict whether a given individual will conduct another crime or not, based on his/her criminal behavior history. He is using a basic Classification and Regression Tree (CART) algorithm, which achieves 74.56% ROC AUC on the training set and only 59.83% ROC AUC on the test set. What would you advice Mr. Red to do to improve his model performance? Please, provide at least 3 recommendations with grounded explanation.

1.I would advise to use cross validation to get the parameters which may increase roc\_score if he put scoring-roc\_auc. test on different Test sets.

2.Second, he should check balance of the data, if number of criminal observations is very small in the data he should use state class weight=’balanced’

3. Use multiple trees (random forests) to predict the outcomes. Random forests in general perform well than a single decision tree as they manage to reduce both bias and variance. They are less prone to overfitting as well.

4.Use other models (for ex Logistic regression)