# Project 2

Jake Carlson

October 4, 2017

**Abstract**

Classification on federal payroll data.

# Contents

# 1    Business Understanding

In this project, I will again be analyzing the federal payroll data obtained by BuzzFeed News through the Freedom of Information Act. I will continue my analysis of the presidency of George W. Bush and Barack Obama by creating classification models based on the payroll data. Based on attributes of the employees, I will be trying to predict class labels such as the annual pay of the employee, what their level of education is, how long they have been a federal employee, what their supervisory status is, and where they are located. The various models will provide insight into how relevant each attribute is as predicitng the class label. For example, the heirarchy of splits in decision trees indicate what attributes are most relevant when determining the class label. By looking at how these models differ between the presidents, and how the ranking of attributes changes, we will gain insight into how each administration restructured the composition of the federal government.

I will use a variety of classification models, including Decision Trees, K-Nearest Neighbors, PART, Artificial Neural Networks, and Random Forests. I will compare the performance of these models to each other to see what models have the best performance in terms of both accuracy and the time required to perform classification on test data.

## 1.1    Annual Pay

By predicting Pay, we will see what attributes correspond to a higher pay rate. This would be useful for employees so that they can see what attributes they should look to change about themselves if they are looking to get a pay raise. For example, if Education has a strong relationship with higher pay rates, employees should consider pursuing higher education in order to recieve a raise.

## 1.2    Education Level

By predicting Education, we will be able to see what attributes are most related with having a higher or lower education level. If a particular agency is in need additional specialized labor, they could offer a subset of their employees financial aid to pursue higher education. This would allow that agency to promote from within, rather than looking for new employees.

From Project 1 we determined that the vast majority of employees had either a high school diploma or a Bachelor's degree. This class imbalance will need to be addressed when creating classification models.[1]

## 1.3    Length of Service

By predicting Length of Service, we will see what factors encourage employees to continue their work at the federal government. By examining how the most imporant attributes change between administrations, we will get an idea of what each administration favored in their employees.

3

We will also see what types of employees prefered to continue their employment in response to the leadership change. Useful why??

## 1.4 Supervisory Status

By seeing what factors affect Supervisory Status, we will see what the most common attributes are for members of leadership within an agency. This could be used by each agency to see if their leadership is biased to employees with certian attributes. With this information, agencies could work to diversify their leadership and potentially improve the operations of their agency and improve the propensity of potential employees to apply for employment from that agency. A more diverse applicant pool would allow the agency to restructure more efficiently under new leadership.

## 1.5 Location

By modeling where an employee is located based on their attributes, we can see what the demographics are in each state. By breaking up this model by agency, we can get an idea of how the agency is organized at a national level. We may see more supervisors in the Washington D.C. and Maryland area. With this model, an agency could more easily visualize their structure and, if a certain state is struggling to meet operational requirements, allocate additional supervisors from a state with a surplus to the state that is struggling.

Because larger states have more federal employees, we will have a class imbalance where large states are over-represented.

# 2 Data Preparation

To prepare my data for classification, I reprocess the raw payroll data. I start by removing columns that I don't plan on using for classification. The attributes I save are given in Table 1. I then replace unknown values with NA, make Pay a numeric attribute, pull out the encoding for what state the employee works in from Station, and create a column that holds the whole agency name for each employee. I join the four quarters for each of the four years I'm looking at into one data frame, and save this frame to disk.

When I load these data frames, I run some additional preparation to make sure the data is ready for classification and make sure the models will train as efficiently as possible. I create a region column which translates the state encoding in Station to the actual name of the state. I convert Pay to an ordinal variable from a continuous variable by descretizing into six classes. The Pay groups are given in Table 2. I then convert the ordinal Age variable to a ratio variable by taking the middle of the age range for each occurance The Age translation is given in Table 3. I break Education up into groups by what level of education the employee achieved. The groups I used are listed in Table 4. Next I convert the ordinal Length of Service variable to a ratio varible by taking the middle of the time span for each occurance. The translation for LOS is given in Table 5. My

| Attribute | Scale | Description |
|---|---|---|
| Agency | Categorical | The four character encoding for the agency the employee w |
| Station | Categorical | The state the employee works in. |
| Age | Ordinal | The age range of the employee, given in 5 year inter |
| Education | Ordinal | The education level achieved by the employee. |
| LOS | Ordinal | The number of years the employee has worked in the federal governemnt, |
| Category | Categorical | The general type of work the employee does, following the PA |
| Pay | Ratio | The annual pay the employee receives in U.S. Dolla |
| SupervisoryStatus | Categorical | The level of leadership the employee has achieved |

Table 1: Attributes To Be Used For Classification

| Pay Ranges |
|---|
| <30k |
| 30-50k |
| 50-70k |
| 70-90k |
| 90-110k |
| >110k |

Table 2: The Pay Ranges Used To Descretize Pay

preparation function also allows for a subset of agencies to be choosen
from the saved data files. The last step in the preparation function is
converting the agencies to a one-hot encoded representation. This creates
a new column for each agency which has a 1 in it if the employee works
at that agency, and a 0 if the employee does not work at that agency.

The scale and range of all of the attributes in the final data set are
given in Table 6. When I go to create a model based on a subset of my
choosen columns, I will throw out any entries that are incomplete. The
count of the number of records that are complete, as well as the percentage
of records that are complete for each year are given in Table 7.

# 3   Modeling

Next I will begin creating models for the each of the cases outlined in the
Business Understanding. I will create two models for each of the cases

| Age Range | Ratio Value |
|:---------:|:-----------:|
| 15-19 | 17 |
| 20-24 | 22 |
| 25-29 | 27 |
| 30-34 | 32 |
| 35-39 | 37 |
| 40-44 | 42 |
| 45-49 | 47 |
| 50-54 | 52 |
| 55-59 | 57 |
| 60-64 | 62 |
| 65-69 | 67 |
| 70-74 | 72 |
| 75+ | 75 |

Table 3: Original Age Ranges In Years And The Choosen Value For Classification

| Group | Education Levels | Description |
|:-----:|:----------------:|:-----------:|
| Elm | 0, 1 | Reached or completed elementary school |
| HS | 3, 4, 5, 6 | Reached or completed high school or an occupational program |
| Col | 7, 8, 9, 10, 11, 12, 13 | Reached or completed college with a Bachelor's degree |
| Grad | 14, 15, 16, 17, 18, 19, 20 | Any level of graduate studies, excluding a Doctorate |
| Doc | 21, 22 | A Doctorate or Post-Doctorate degree |

Table 4: Ordinal Education Groups

| LOS Range | Ratio Value |
|:---------:|:-----------:|
| < 1 | 1 |
| 1-2 | 1 |
| 3-4 | 3 |
| 5-9 | 7 |
| 10-14 | 12 |
| 15-19 | 17 |
| 20-24 | 22 |
| 25-29 | 27 |
| 30-34 | 32 |
| 35+ | 35 |

Table 5: Original Length Of Service Ranges And The Choosen Value For Classification

| Attribute | Scale | Range |
|---|---|---|
| Agency | Categorical | Unique four characters for each agency |
| Station | Categorical | 1-56, some values missing |
| Age | Ratio | 17-72 by 5, 75 |
| Education | Categorical | 1-22 |
| LOS | Ratio | 1, 3, 7-32 by 5, 35 |
| Category | Categorical | P, A, T, C, O, B |
| Pay | Ordinal | <30k, 30-50k, 50-70k, 70-90k, 90-110k, >110k |
| SupervisoryStatus | Categorical | 2, 4, 5, 6, 7, 8 |

Table 6: Final Data Set Attributes

| Year | Number Complete | Percentage of Original |
|---|---|---|
| 2001 | 2,535,278 | 58.09% |
| 2005 | 2,598,800 | 55.05% |
| 2009 | 2,862,972 | 55.22% |
| 2013 | 3,136,697 | 58.91% |

Table 7: The Number Of Complete Records For Each Year

and compare their performance using 10-fold cross validation. I will select the model that perfoms the best and train it on the data for each of my choosen years. I will then compare these models to see how their structure changes with the presidency. To make training and evaluation easier, I will create a pipeline to break up the data, train the models, test them, and return the results.

## 3.1 Annual Pay

I will be modeling Pay for the Department of Justice. I will compare the performace of decision trees and random forests. In order to speed up the training of my models, I only want to pass in the attributes that are most important for predicting Pay. To determine these attributes I examine the results of both cfs, which uses correlation and cross entropy, and consistency, which uses a consistency measure, to determine variable importance. The most important attributes given by consistency in descending order are Age, Education, Length of Service, Category, and SupervisoryStatus. The most important measures by cfs are Education and Category. I will use Age, Education, and Category for training my models.

The accuracy and kappa scores for each of the models is given in Table 8. These scores aren't great, but the random forests consistently beat out the decision tree. The most important variables given by the random forests are Education and Age followed by whether or not the employee is in a technical role or a clerical role. The decision tree concurs with the random forests that Education is the most important attribute, followed by whether the employee is in a techinical role or not. This indicates that the education is strongly related with how much an employee makes in

7

| Accuracy | | | | | | | |
|---|---|---|---|---|---|---|---|
| Classifier | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| Decision Trees | 0.5176392 | 0.5211121 | 0.5275774 | 0.5289369 | 0.5376409 | 0.5415454 | 0 |
| Random Forests | 0.5619261 | 0.5658214 | 0.5704048 | 0.5698463 | 0.5739125 | 0.5764363 | 0 |
| Kappa | | | | | | | |
| Classifier | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| Decision Trees | 0.3466076 | 0.3497777 | 0.3672808 | 0.3688381 | 0.3881202 | 0.3946472 | 0 |
| Random Forests | 0.4274982 | 0.4319667 | 0.4382973 | 0.4374711 | 0.4428353 | 0.4456237 | 0 |

Table 8: The Accuracy And Kappa Score for Decision Trees and Random Forests

| | 2005 | 2005 | 2013 | 2013 |
|---|---|---|---|---|
| Decision Tree | Education | 9932.9 | Education | 9382.2 |
| | CategoryT | 7682.5 | Age | 7867.3 |
| | CategoryP | 5599.9 | CategoryP | 5382.3 |
| | CategotyC | 5305.0 | CategoryB | 4715.5 |
| Random Forest | Education | 10345.0 | Education | 11113.4 |
| | Age | 4089.5 | Age | 5470.7 |
| | CategoryT | 2805.7 | CategoryC | 1601.5 |
| | CategoryC | 2393.5 | CategoryT | 1476.4 |

Table 9: Variable Importance Rankings for Predicting Pay for Decision Trees and Random Forests in 2005 and 2013

the Department of Justice.

The structure of the decision tree given in Figure 1 shows that clerical roles tend to be paid less and technical roles tend to be paid more. I want to see how these models change under the Obama administration, so I will rerun these using the 2013 data set. Again, Education is the most important attribute for both the decision tree and the random forest, followed by Age. The structure of the decision tree for 2013 is given in Figure 2. We see that the splits in the decision tree closest to the root have to do with what Category an employee is in. The difference in the complexity of the decision tree would appear to indicate that, under the Obama administration, pay was more closely related to what type of position you held.

## 3.2   Education Level

I will model Education for the Department of Justice. The models for Pay showed that the Category and Education are the most relevant attributes
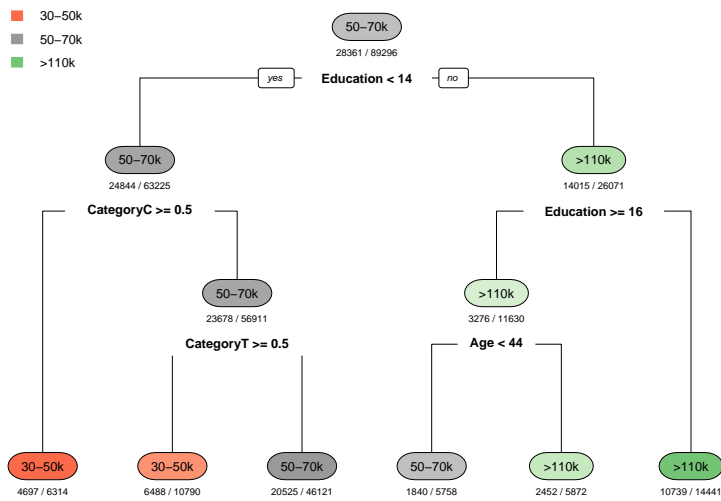
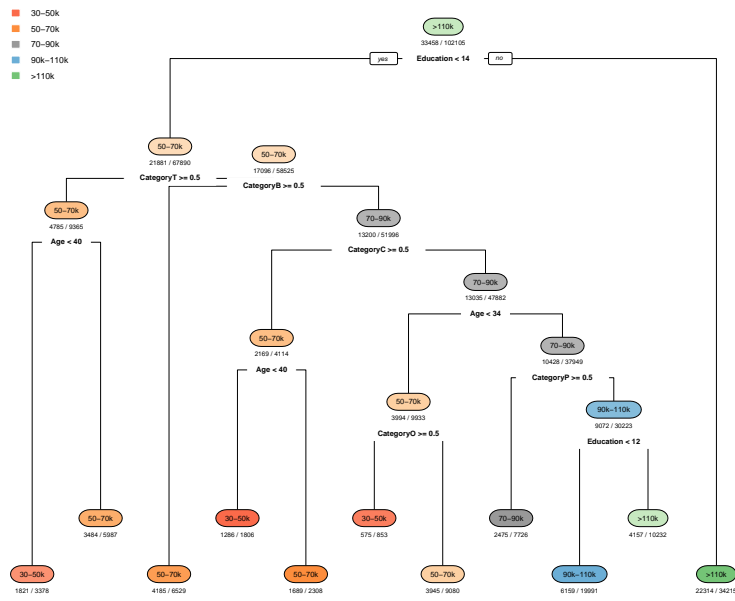Figure 1: The Structue of the Decision Tree for Pay in the Department of Justice in 2005

Figure 2: The Structue of the Decision Tree for Pay in the Department of Justice in 2013

|                | 2005       | 2005    | 2013       | 2013    |
|----------------|------------|---------|------------|---------|
| Decision Tree  | Pay.110k   | 11733.8 | Pay.110k   | 11483.9 |
|                | CategoryP  | 9518.1  | CategoryP  | 10257.4 |
|                | Pay30.50k  | 3680.3  | Pay50.70k  | 5405.1  |
|                | CategoryB  | 2814.5  | CategoryB  | 3793.9  |
| Random Forest  | CategoryP  | 9515.0  | CategoryP  | 10257.5 |
|                | Pay.110k   | 3171.8  | Pay.110k   | 3554.9  |
|                | Age        | 930.5   | CategoryB  | 1421.8  |
|                | CategoryB  | 915.6   | Age        | 1384.9  |

Table 10: Variable Importance Rankings for Predicting Education for Decision Trees and Random Forests in 2005 and 2013

for predicting the Pay of an employee. I would expect Category to be an important attribute in predicting Education. I will again use cfs and consistency to determine the most relevant attributes to use when training my models for Education. For this attribute, I will again compare random forests to decision trees. Consistency gives the most important attributes as Age, Length of Service, Category, and SupervisoryStatus. Cfs gives the most important attributes as Category, and Pay. I will use Age, Category, and Pay to build my models.

The models had comparable accuracy and kappa scores to the scores in Table 8. Again, the random forest consistently beat the decision tree with a maximum accuracy of 0.616 over 0.604 and a maximum kappa of 0.415 over 0.399. The most important attributes for 2005 and 2013 are given in Table 10. I've also provided the decision tree structure for 2005 in Figure ?? and 2013 in Figure 4.

It is interesting how similar the structures of the trees are. The trees are identical except for the last split between college and graduate education levels. The relationship between reaching graduate studies and having a pay greater than 100k. The most accurate class in both decision trees is where the employee reached graduate level studies and has an annual income greater than 100k. The accuracy of this class actually drops moving from Bush to Obama, indicating employees with lower levels of education were able to reach a higher pay grade.

## 3.3   Length of Service

## 3.4   Supervisory Status

## 3.5   Location

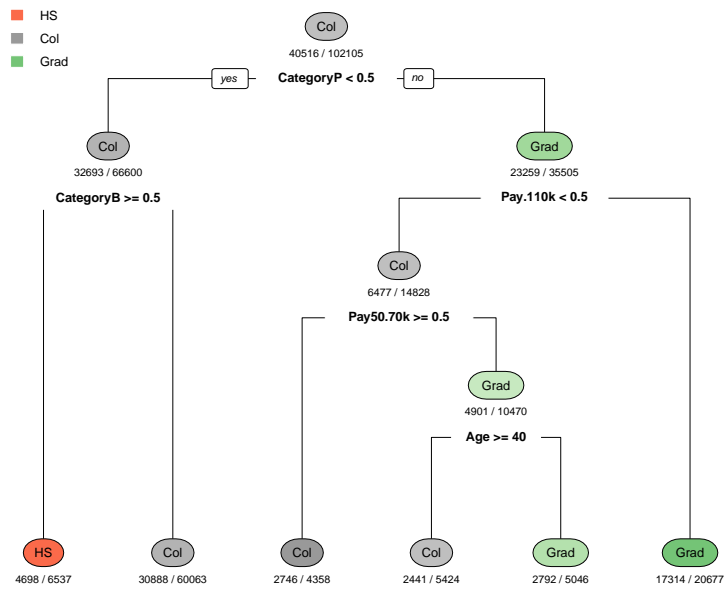What are the differences across time and between agencies?

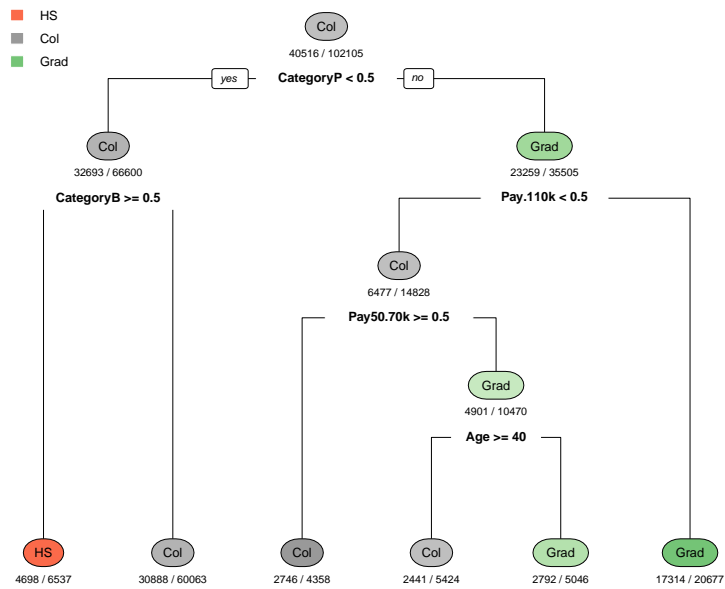Figure 3: The Structue of the Decision Tree for Education in the Department of Justice in 2013

Figure 4: The Structue of the Decision Tree for Education in the Department of Justice in 2013

# 4    Evaluation and Deployment

Are these models useful for solving the problems outlined in the business understanding?

# References

[1] Jake Carlson *CSE 5331 - Data Mining Project 1*
https://github.com/jakecarlson1/data-mining-projects/blob/master/
project-1/report/carlson-project-1.pdf