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, Artificial Neural Networks, and Random Forests. I will compare the performace 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.

We will also see what types of employees prefered to continue their employement 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 four 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 convert Education to an integer to improve the training speed with this ordinal variable. 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 4. My preparation function also

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
<50k
50-75k
75-100k
>100k

Table 2: The Pay Ranges Used To Descretize Pay

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.

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 5.

3 Modeling

5 classification methods, advantages and disadvantages, most important features, performance.

What are the differences across time and between agencies?

4 Evaluation and Deployment

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

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

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

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 5: The Number Of Complete Records For Each Year

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