

Project 2

Jake Carlson

October 4, 2017

Abstract

Classification on federal payroll data.

Contents

1	Business Understanding	3
1.1	Annual Pay	3
1.2	Education Level	3
1.3	Length of Service	3
1.4	Supervisory Status	4
1.5	Location	4
2	Data Preparation	4
3	Modeling	4
4	Evaluation and Deployment	4

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 predicting the class label. For example, the hierarchy 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 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 receive 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 important 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 preferred 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 certain 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

Cleaning, features removed/created.

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?

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>