Predictive Model for Federal Wages

Jenn Le

10/24/2017

Abstract

Some stuff for an abstract here

Contents

Business Understanding

Data Preparation

Cleaning

For this project, I started with uncleaned data from the non-department of defense data from the years 2001, 2005, 2009, and 2013. Since there is so much data, I decided that removing any NA's or unknowns would be more beneficial to the predictive model than imputing data. I chose to leave duplicate ID's alone because they could indicate a pay raise based on a new degree or change of agency. I then removed the psuedo ID and name features since they are not meaningful to the class that I am trying to predict. To make the data more meaningful at a glance, I also chose to replace encoded nominal attributes in the data with their actual values such as replacing the station codes with the cooresponding states. However, since decision trees are slow with ordinal and nominal attributes that have a lot of possible values, I converted the date, age, education, length of service, supervisory status, appointment and NSFTP into continuous values. I also removed any agencies that had less than 10,000 members so I could focus on larger agencies. After doing all of this, there are 12801507 records left from the original 19645240 which is about 65.16%.

To create the classes, I rounded up pay to the closest multiple of $25,000 and treated any salaries above $200,000 to be the same. I decided to do this in order to retain the meaning that ordinals have in relation to one another while encapsulating a range of values. I chose to leave the date attribute in rather than correcting for inflation to see if my feature selection method would pick up a strong relationship between date and pay. In addition, I renamed the features to make them more consistent with the code.

Feature Selection

I used the consistency method from FSelector to select a subset of the features. The ones selected were age, education, length of service, category, and supervisory status. However, I also chose to keep agency and state because those are important choices people make when choosing to work for the government.

Final Dataset

Feature	Scale	Description
Agency	Nominal	The name of the agency the employee works for
State	Nominal	The state the employee works in
Age	Ordinal	The age range the employee belongs to
Education	Ordinal	Encoded education level of the employee
Length of Service	Ordinal	Number of years the employed
Category	Nominal	Type of work the employee does
Supervisory Status	Nominal	Employee's authority

Table shows the features that are used in this project along with their scales and descriptions.

Modeling

Splitting Data

I first graphed the class counts to determine if there is a class imbalance.

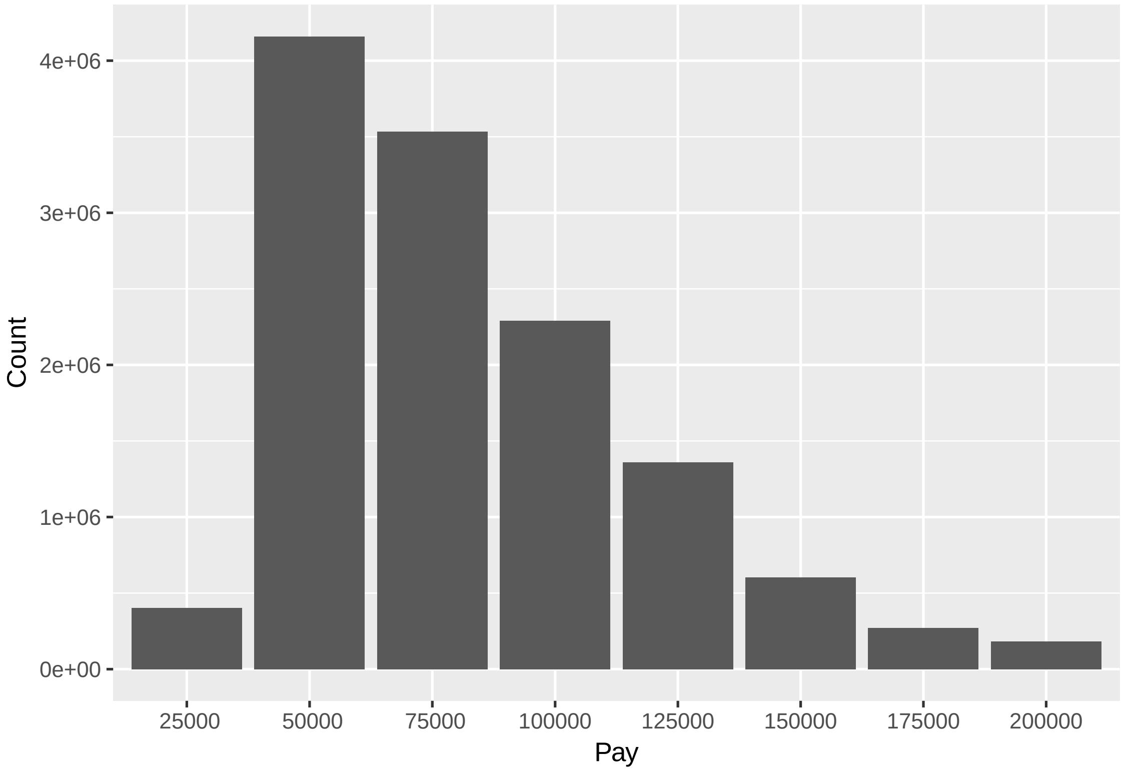
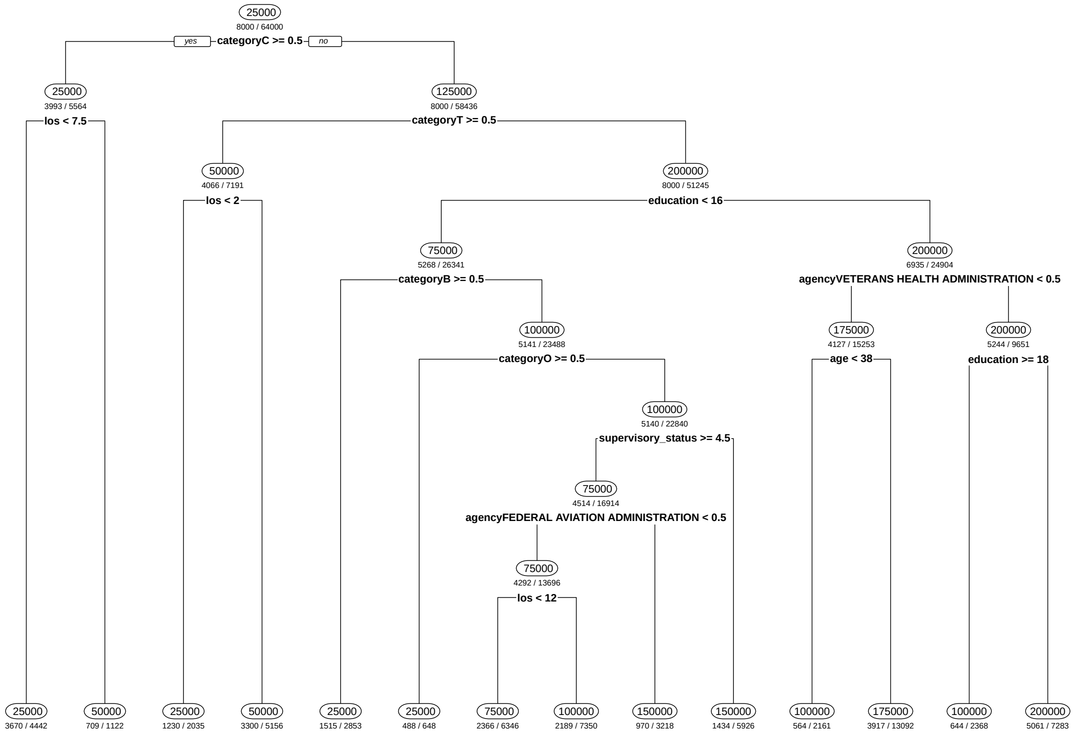


Figure 1 shows that the data is

I decided to balance the classes and use 20% holdout to split the data into training and testing sets. For training, I created splits using 10-fold cross validation and used these splits for each of the models I trained in order to stay consistent.

Price CART Model

cp	Accuracy	Kappa	AccuracySD	KappaSD
0.0037321	0.4308438	0.3495357	0.0099514	0.0113730
0.0046964	0.4249375	0.3427857	0.0152723	0.0174540
0.0063214	0.4204062	0.3376071	0.0142235	0.0162554
0.0068929	0.4165781	0.3332321	0.0114207	0.0130523
0.0082321	0.4080938	0.3235357	0.0127089	0.0145245
0.0082857	0.4067344	0.3219821	0.0124612	0.0142414
0.0115625	0.3741406	0.2847321	0.0299249	0.0341999
0.0247857	0.3535156	0.2611607	0.0240436	0.0274784
0.0435000	0.3323750	0.2370000	0.0167708	0.0191666
0.0729881	0.2231406	0.1121607	0.0781970	0.0893680



Evaluation and Deployment

My models would be useful for anyone looking to get a job with the federal government. They can see how much they can expect to receive in compensation in certain departments and states based on their education level.