Predictive Model for Federal Wages

Jenn Le

10/24/2017

Abstract

This investigation examines several CART models built to perform classification of federal employment data. The models aim to predict employee salaries, education, and supervisory statuses for the years 2001, 20015, 2009, and 2013 with a particular emphasis on 2005 and 2013 when predicting pay. Feature importance is examined for each model to determine the factors that affect an employee's pay in an attempt to help prospective employees make decisions about their career path. Classification of this highly varied data is found to be inaccurate when using decision trees but close enough for this business case to be useful.

Contents

Business Understanding

Observing and predicting trends in government employment data is important to both prospective government employees as well as the government themselves. Prospective employees that know what to expect in terms of compensation according to their credentials and the job conditions can make more informed choices when deciding where to work or whether or not to get a more advanced degree. The government can make use of these models to figure out if an employee is being paid too little or too much compared to others in the same situation. This can results in better informed raises or offers for new employees. Looking at the factors that affect employee salaries can also help to give a better understanding of the differences in presidential terms.

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 and states that had less than 10,000 members so I could focus on environments people are more likely to pursue a career in. 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

Table 1: Dataset Features

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 1 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.

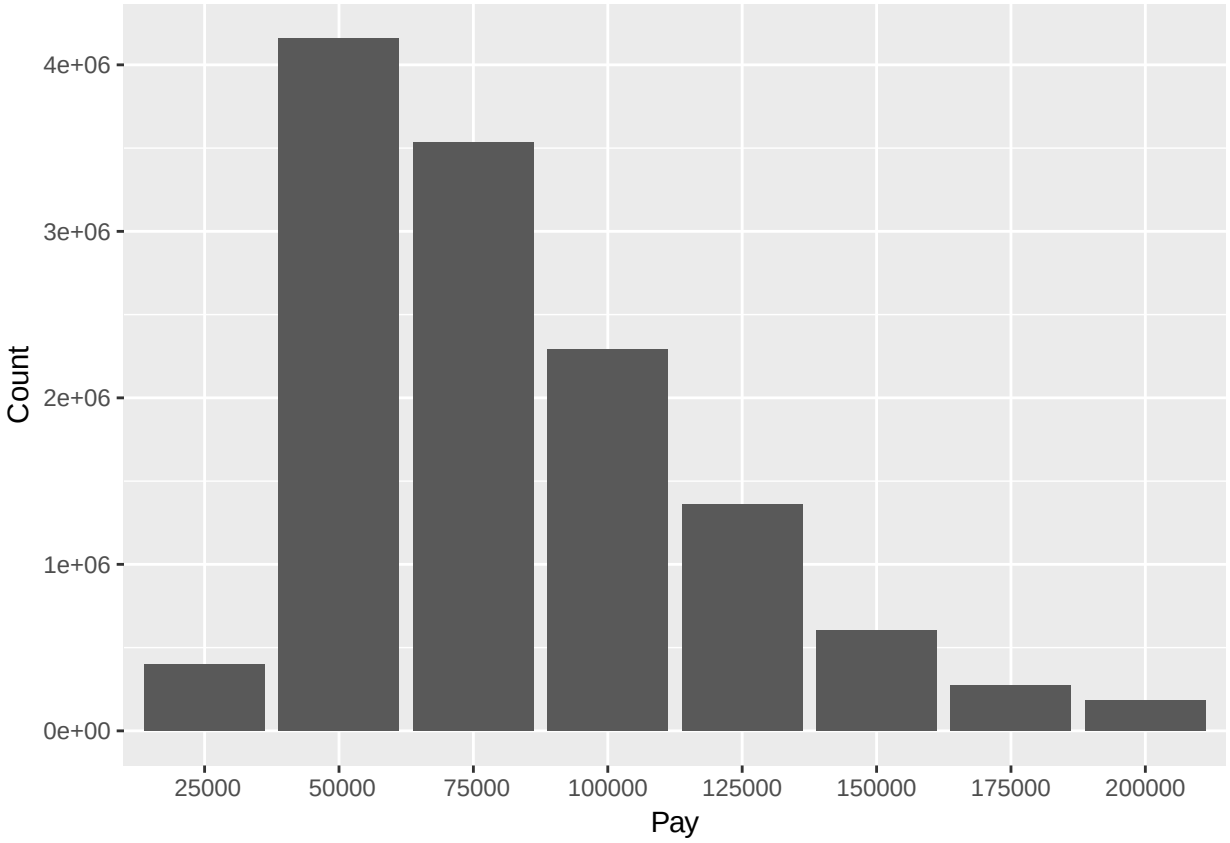

Figure 1: Class Counts

Figure 1 shows that the data is unimodal and skewed right. The most instances lie between $25,000 and $50,000 and decrease as the range gets larger. With the class distribution like this, it is likely that any models created will completely ignore the classes 25000, 150000, 175000, and 200000.

I decided to balance the classes by downsampling and then using 20% holdout to split the balanced data into training and testing sets. For training, I used the 10-fold cross validation built into caret.

Price Prediction

General

The first prediction model I created was trained on data from 2001, 2005, 2009, and 2013 and predicts the pay range of an employee based on the features listed in table 1. The data was first balanced with 10,000 instances of each class. Using a decision tree for this classification allows for transparency in our model which helps us understand how each factor affects the pay of an employee. In this model, there are two features with many possible values, state and agency, which can be examined closely using a decision tree. This information would be lost in a more black box approach.

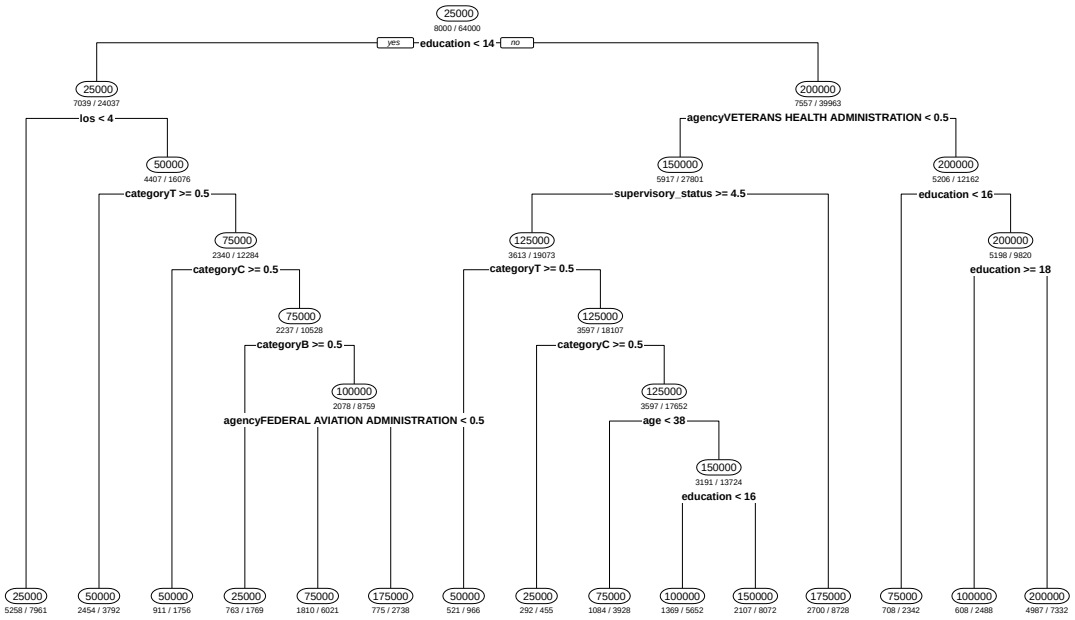

Figure 2: Pay Prediction CART Model

Figure 2 shows the decision tree generated by the model. From this tree, we can see that this model deems education to be the most important factor in an employee's pay. The cutoff for education here is a master's degree so according to this model, getting at least a master's degree would be more beneficial to a prospective employee than going to work sooner. It also shows, however, that an employee with a lower education level can hope to make a lot of money depending on the type of work they do with a large pay increase if they work in the Federal Aviation Administration. This detailed model shows us that the agencies that have the highest affect on pay are the Federal Aviation Administration and the Veteran's Health Administration. Prospective employees with a lower education level should aim to work in the Federal Aviation Administration while ones with a higher education level should aim for the Veteran's Health Administration.

Table 2: Pay Prediction Results

Accuracy	0.4138125
Kappa	0.3300714
AccuracyLower	0.4061660
AccuracyUpper	0.4214903
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

Table 3: Pay Prediction Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1608	613	222	86	45	22	8	5
50000	288	988	320	23	12	4	0	0
75000	70	308	893	829	478	274	126	49
100000	27	60	357	484	458	345	209	117
125000	0	0	0	0	0	0	0	0
150000	7	7	110	270	401	543	473	155
175000	0	14	89	272	439	750	881	450
200000	0	10	9	36	167	62	303	1224

Table 2 shows that the accuracy of this model on the cooresponding training data is much less than desired with an even lower kappa statistic. This is expected, however, since the data I am working with is much too complicated for a decision tree to perform well. Table 3 shows that while the predictions are inaccurate, they are not too far off from their actual values. Because our business case is more focused on a general idea of whether or not a certain change in a factor will increase pay, the performance of this model is perfectly acceptable.

Table 4: Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.8040	0.9285000	0.6163281	0.9707266	0.6163281	0.8040
Class: 50000	0.4940	0.9537857	0.6042813	0.9295510	0.6042813	0.4940
Class: 75000	0.4465	0.8475714	0.2950116	0.9146689	0.2950116	0.4465
Class: 100000	0.2420	0.8876429	0.2352941	0.8912716	0.2352941	0.2420
Class: 125000	0.0000	1.0000000	NaN	0.8750000	NA	0.0000
Class: 150000	0.2715	0.8983571	0.2761953	0.8961807	0.2761953	0.2715
Class: 175000	0.4405	0.8561429	0.3043178	0.9146127	0.3043178	0.4405
Class: 200000	0.6120	0.9580714	0.6758697	0.9453097	0.6758697	0.6120

Table 4 shows detailed statistics on each of the classes for this prediction model. One point that stands out is that 125000 has a sensitivity of 0 and a specificity of 1 because the model did not predict a single instance of this class.

President Bush

I trained the CART model on data specifically from the year 2005 to see how the same features had affected pay during President Bush's term in office, especially in comparison to President Obama's term.

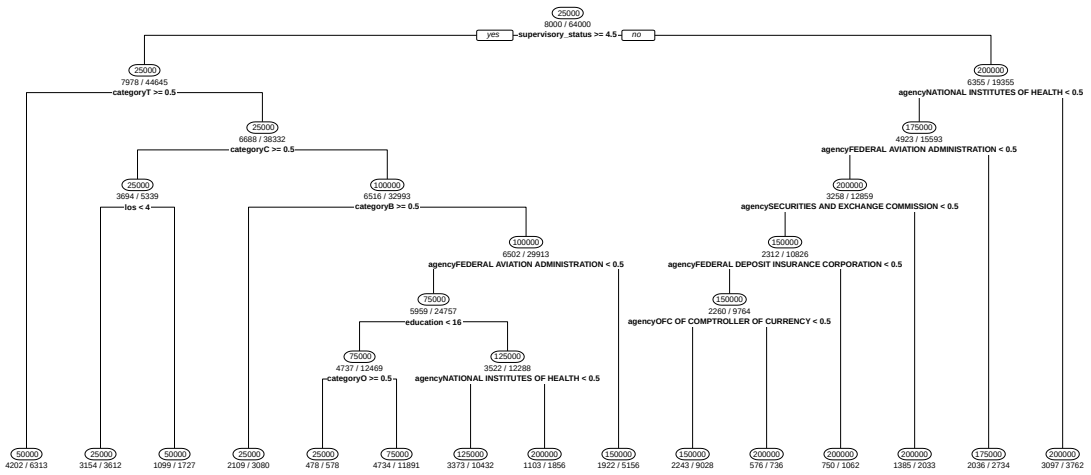

Figure 3: Pay Prediction 2005 CART Model

Figure 3 shows that supervisory status is the most important feature in this model. Higher authority can expect to make much more in most cases. With a lower level of authority, however, working for the Federal Aviation Administration or having a higher level of education is then the best bet for a higher salary.

Table 5: Pay Prediction 2005 Results

Accuracy	0.4983125
Kappa	0.4266429
AccuracyLower	0.4905348

AccuracyUpper	0.5060909
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	NaN

Table 6: Pay Prediction 2005 Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1438	347	64	7	0	0	0	0
50000	463	1304	215	20	2	0	0	0
75000	39	239	1168	941	371	140	55	41
100000	0	0	0	0	0	0	0	0
125000	47	36	276	518	819	590	180	81
150000	11	73	254	468	700	1036	885	135
175000	0	0	0	6	36	118	465	0
200000	2	1	23	40	72	116	415	1743

Table 5 shows that the kappa statistic for this model is higher than the model shown in table 2. This could probably be attributed to the fact that the supervisory status in this model is more indicative of an employee's pay range and only further takes into account which agency they work in. On the other hand, the previous model includes a lot more choices that are also more varied. The two models are similar in that their predictions are typically not far off from the actual values.

Table 7: Pay Prediction 2005 Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.7190	0.9701429	0.7747845	0.9602658	0.7747845	0.7190
Class: 50000	0.6520	0.9500000	0.6506986	0.9502715	0.6506986	0.6520
Class: 75000	0.5840	0.8695714	0.3901136	0.9360295	0.3901136	0.5840
Class: 100000	0.0000	1.0000000	NaN	0.8750000	NA	0.0000
Class: 125000	0.4095	0.8765714	0.3215548	0.9122129	0.3215548	0.4095
Class: 150000	0.5180	0.8195714	0.2908478	0.9224956	0.2908478	0.5180
Class: 175000	0.2325	0.9885714	0.7440000	0.9001626	0.7440000	0.2325
Class: 200000	0.8715	0.9522143	0.7226368	0.9810863	0.7226368	0.8715

Table 6 shows that this training set did not predict any values of 100000. This could be because the model is being trained to skew towards more extreme values and as a result, always choose a lower or higher range rather than the neutral one.

President Obama

I trained data exclusively from 2013 under the same conditions as the two previous models in order to draw a comparison between the three and examine how the factors that affect pay change across presidential terms. The transparency of decision trees allows us to look at how similarly decisions are made and make comparisons about the overall structure of the model.

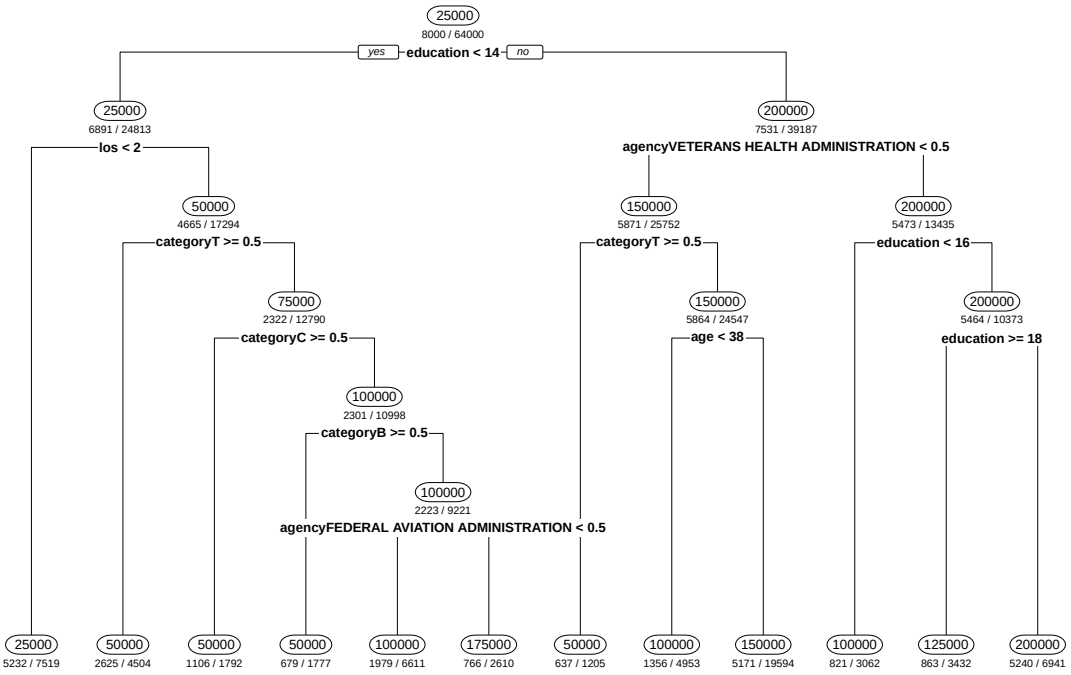

Figure 4: Pay Prediction 2013 CART Model

Figure 4 shows that this model follows a similar pattern to the general pay prediction model, shown in figure 2. Education is the most important feature, with the cutoff at the master's degree level. The difference is that this model does not place an emphasis on supervisory status. After education, the most important features are the length of service as well as the type of work being done.

Table 8: Pay Prediction 2013 Results

Accuracy	0.4141875
Kappa	0.3305000
AccuracyLower	0.4065399
AccuracyUpper	0.4218663
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

Table 9: Pay Prediction 2013 Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1312	414	99	40	21	11	5	4
50000	374	1253	608	43	14	3	3	0
75000	0	0	0	0	0	0	0	0
100000	192	228	902	1029	670	387	172	64
125000	62	31	103	206	238	105	53	56
150000	59	68	253	607	909	1286	1258	488
175000	0	0	26	62	112	165	222	101
200000	1	6	9	13	36	43	287	1287

Table 8 shows that the accuracy and kappa score of this model are similar to those of the first model. This makes sense seeing as how the structure of both decision trees are also very similar to one another.

Table 10: Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.6560	0.9575714	0.6883526	0.9511849	0.6883526	0.6560
Class: 50000	0.6265	0.9253571	0.5452567	0.9454824	0.5452567	0.6265
Class: 75000	0.0000	1.0000000	NaN	0.8750000	NA	0.0000
Class: 100000	0.5145	0.8132143	0.2823820	0.9214147	0.2823820	0.5145
Class: 125000	0.1190	0.9560000	0.2786885	0.8836657	0.2786885	0.1190
Class: 150000	0.6430	0.7398571	0.2609578	0.9355130	0.2609578	0.6430
Class: 175000	0.1110	0.9667143	0.3226744	0.8838819	0.3226744	0.1110
Class: 200000	0.6435	0.9717857	0.7651605	0.9502025	0.7651605	0.6435

Table 10 shows that no instances of 75000 were predicted this time. The fact that the class that gets ignored is decreasing each time is suspicious and with more time, I would investigate this further.

Comparison

Table 11: Pay Prediction Variable Importance Comparison

Rank	General	Bush	Obama
1	Education (100.00)	Supervisory Service (100.00)	Education (100.00)
2	Length of Service (40.79)	Category C (87.96)	Length of Service (43.93)
3	Veteran's Health Administration (28.50)	Category T (76.19)	Category T (27.72)
4	Category T (27.01)	National Institutes of Health (72.67)	Veteran's Health Administration (26.76)
5	Category C (19.58)	Category B (60.05)	Category C (13.93)

Table 11 shows that the general pay prediction model I trained is much more similar to the model trained with Obama's term than Bush's. This could be because the random sampling done for the general model picked up more instances that fell under Obama's jurisdiction. The model trained with 2005 data finds supervisory status and category to be the most important features while the model trained with 2013 data finds education and length of service to be the most important. It seems like during Bush's term, pay was dictated by the authority you have and the type of work you do. On the other hand, qualification and experience dictated the pay during Obama's term. This could be due to a change in values between the presidents but I feel that it's much more likely to be attributed to a demographic shift that happened in the U.S. in general where more and more people were getting bachelor's degrees so more and more importance was being put upon attaining even higher levels of education.

In each of these pay prediction models, state has not been a factor at all. It turns out that this is because the government has a fixed pay scheme so that location does not affect pay. Knowing this information could help a prospective employee make the decision about where to work. It is important to know because making $100,000 in Texas goes much further than making the same amount of money in a state with a higher cost of living such as California or New York.

Minimized Price Prediction

The feature selection I did using FSelector did not include agency or state so I decided to create a pay prediction model leaving out these features to see if they hindered my original model in any way. Using

fewer features could help the decision tree make more definite splits. The fact that the two features that included a lot of possible values have been removed also means a faster training time for this model as well as a chance to get a clearer understanding of the relationships between the remaining features. However, the models shown in figures 2 and 4 have shown that some splits depend entirely on agency so the accuracy of this model may be negatively impacted by the removal of the feature.

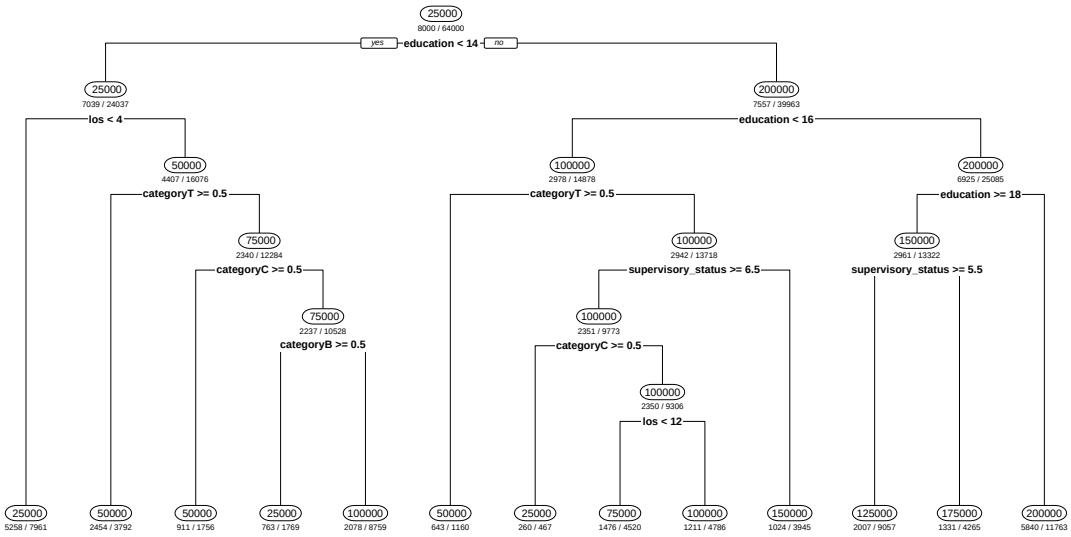

Figure 5: Minimized Pay Prediction CART Model

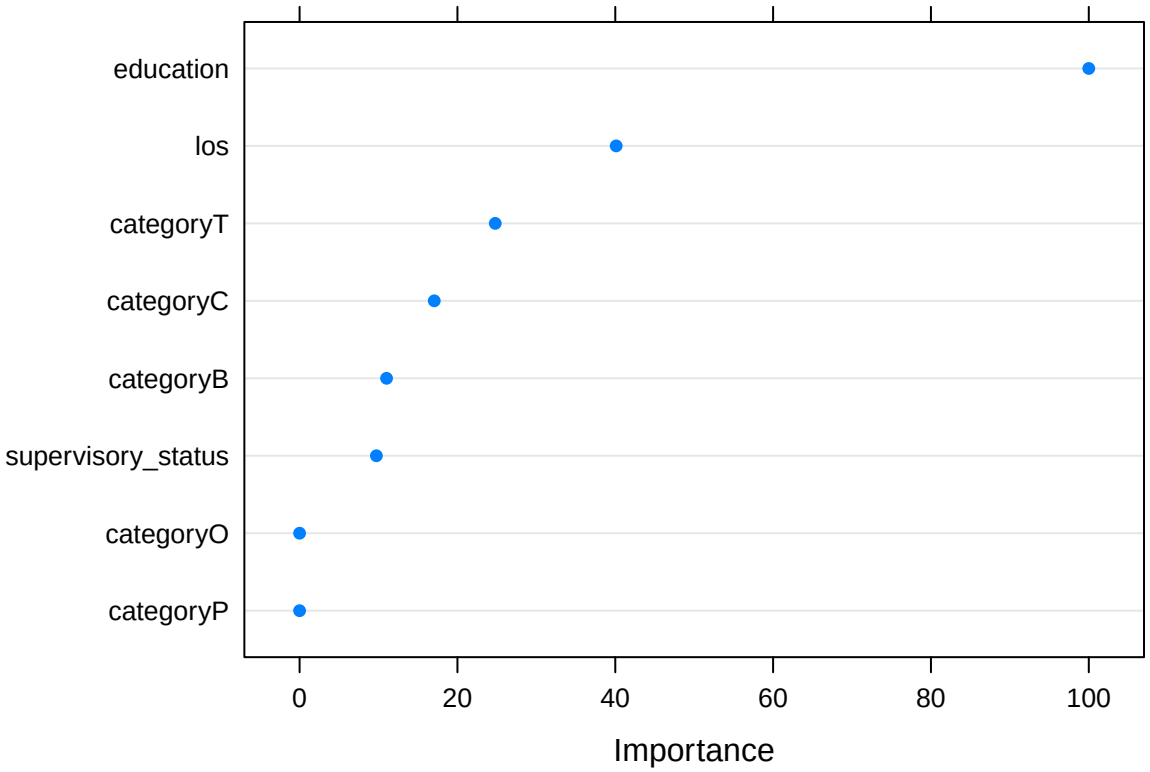

Figure 6: Minimized Pay Prediction Variable Importance

Figures 5 and 6 show that education is the most important factor by far in this model. However,

figure 5 also shows that at lower levels of education, length of service and the type of work being done have more of an impact while at higher levels of education, supervisory status has more of an impact.

Table 12: Minimized Pay Prediction Results

Accuracy	0.3907500
Kappa	0.3037143
AccuracyLower	0.3831789
AccuracyUpper	0.3983608
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

Table 12 shows that the accuracy score and kappa statistic of this model is lower than the first but not by very much at all. In addition, the training time of this is faster so in the case where agency doesn't matter too much, it could be more beneficial to use this model.

Table 13: Minimized Pay Prediction Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1603	630	222	86	45	22	8	5
50000	288	1015	322	25	11	4	0	0
75000	53	146	342	290	155	98	39	15
100000	2	123	699	784	663	579	412	155
125000	52	65	300	480	488	453	279	123
150000	0	5	59	144	187	233	212	109
175000	0	2	23	89	150	281	362	168
200000	2	14	33	102	301	330	688	1425

Table 14: Minimized Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.8015	0.9272857	0.6115986	0.9703266	0.6115986	0.8015
Class: 50000	0.5075	0.9535714	0.6096096	0.9312871	0.6096096	0.5075
Class: 75000	0.1710	0.9431429	0.3005272	0.8884403	0.3005272	0.1710
Class: 100000	0.3920	0.8119286	0.2294410	0.9033617	0.2294410	0.3920
Class: 125000	0.2440	0.8748571	0.2178571	0.8901163	0.2178571	0.2440
Class: 150000	0.1165	0.9488571	0.2455216	0.8825992	0.2455216	0.1165
Class: 175000	0.1810	0.9490714	0.3367442	0.8902513	0.3367442	0.1810
Class: 200000	0.7125	0.8950000	0.4922280	0.9561236	0.4922280	0.7125

Tables 13 and 14 show that this model follows the same pattern as the previous models. Although the accuracy of the predictions is low, the predictions are typically close enough to the actual values that they would not make much of a difference when making career decisions.

Price Prediction Between Agencies

After looking at the general pay prediction models, I wanted to take a more in depth look at how certain factors affect pay within different agencies. To do so, I trained decision trees on the three biggest agencies in the dataset, Internal Revenue Services, the Social Security Administration, and the Veteran's Health Administration. Each model was trained in the same circumstances as the previous models with 10,000 instances of each class split into training and testing sets.

These models will be useful for employees that already know they want to work at one of these agencies and want to know how they can increase their chances at a higher salary. Without having to look at agency, these models will also be faster to train.

Internal Revenue Services

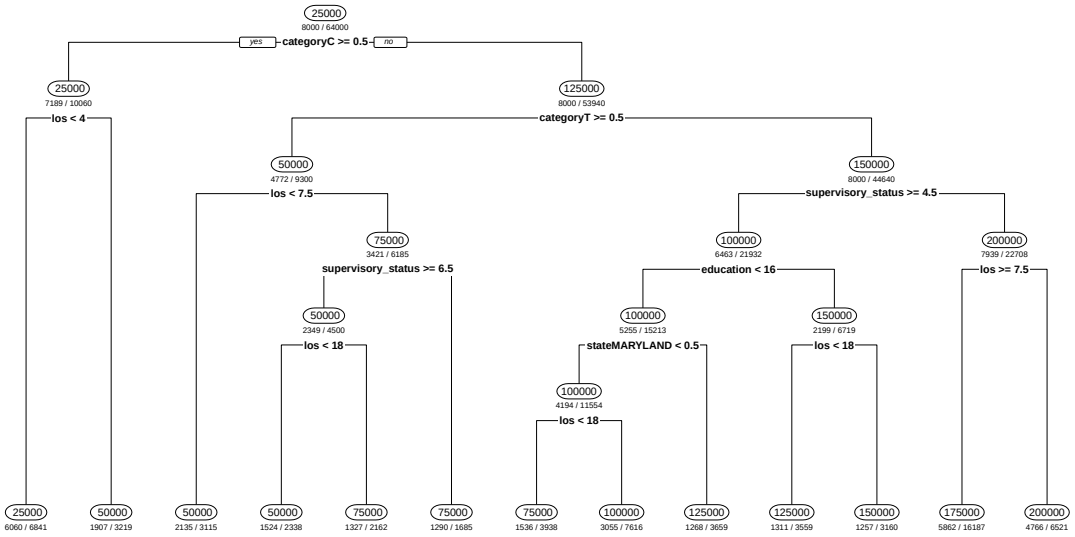

Figure 7: Pay Prediction IRS CART Model

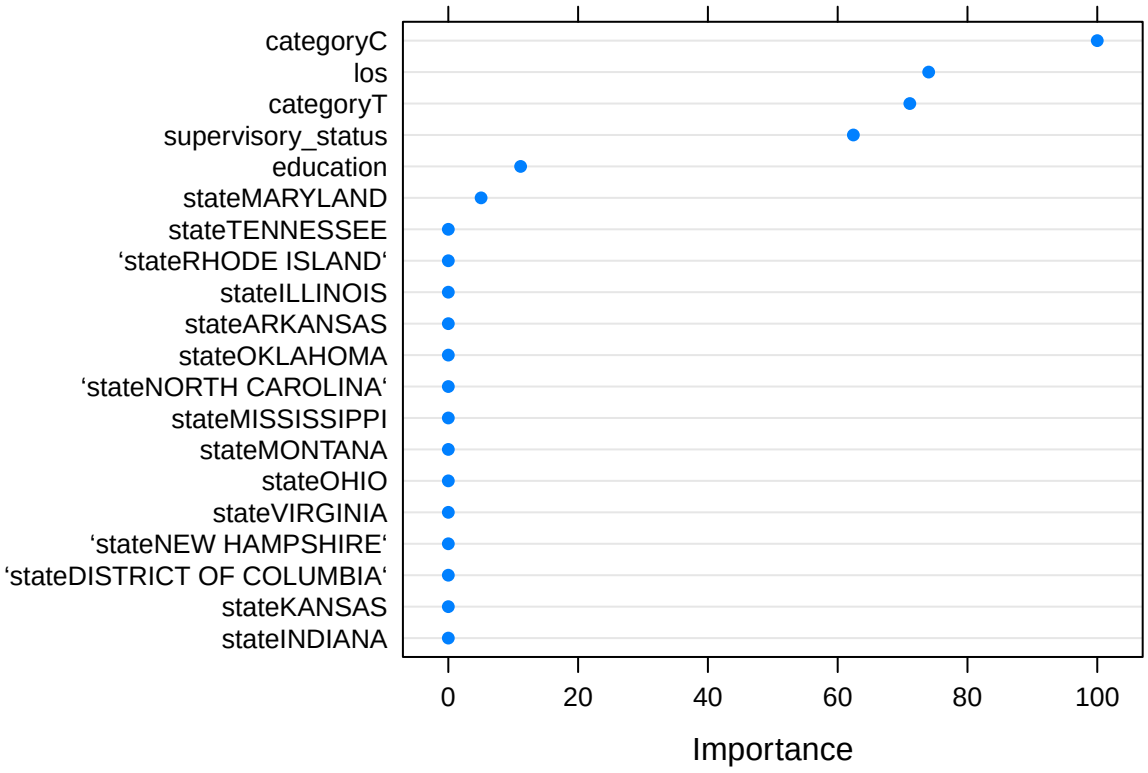

Figure 8: Pay Prediction IRS Variable Importance

Figures 7 and 8 show that the type of work being done and length of service are the most important factors in determining the pay of an IRS employee. Supervisory status is also important but education is not. We can probably assume that this is because employees within the IRS have similar education levels unless they have more authority in which case they have a higher education level so the model views the two factors to be one and the same.

Table 15: IRS Pay Prediction Results

Accuracy	0.5256875
Kappa	0.4579286
AccuracyLower	0.5179150
AccuracyUpper	0.5334507
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	NaN

Tables 15 shows that the accuracy score and kappa statistic of this model is much higher than the previous ones. This is probably because the decision tree can make more definitive choices since it does not have to take agency into account. It is much more accurate than the minimized pay prediction model shown in figure 5 though because it does not completely disregard agency.

Table 16: IRS Pay Prediction Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1521	178	4	0	0	0	0	0
50000	455	1405	330	0	0	0	0	0
75000	19	364	1024	336	131	46	7	6
100000	0	36	435	765	453	166	28	0
125000	5	16	130	497	676	465	66	9
150000	0	1	11	89	165	319	199	0
175000	0	0	62	294	510	878	1490	774
200000	0	0	4	19	65	126	210	1211

Table 17: IRS Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.7605	0.9870000	0.8931298	0.9664965	0.8931298	0.7605
Class: 50000	0.7025	0.9439286	0.6415525	0.9569153	0.6415525	0.7025
Class: 75000	0.5120	0.9350714	0.5297465	0.9306178	0.5297465	0.5120
Class: 100000	0.3825	0.9201429	0.4062666	0.9125168	0.4062666	0.3825
Class: 125000	0.3380	0.9151429	0.3626609	0.9063384	0.3626609	0.3380
Class: 150000	0.1595	0.9667857	0.4068878	0.8895242	0.4068878	0.1595
Class: 175000	0.7450	0.8201429	0.3717565	0.9574716	0.3717565	0.7450
Class: 200000	0.6055	0.9697143	0.7406728	0.9450748	0.7406728	0.6055

Tables 16 and 17 show that there is an even lower distribution range in the model's prediction than with previous models. This makes this model much more useful because there are not as many outliers in the predictions so we can expect the predicted range to be close enough to its actual range to be useful.

Social Security Administration

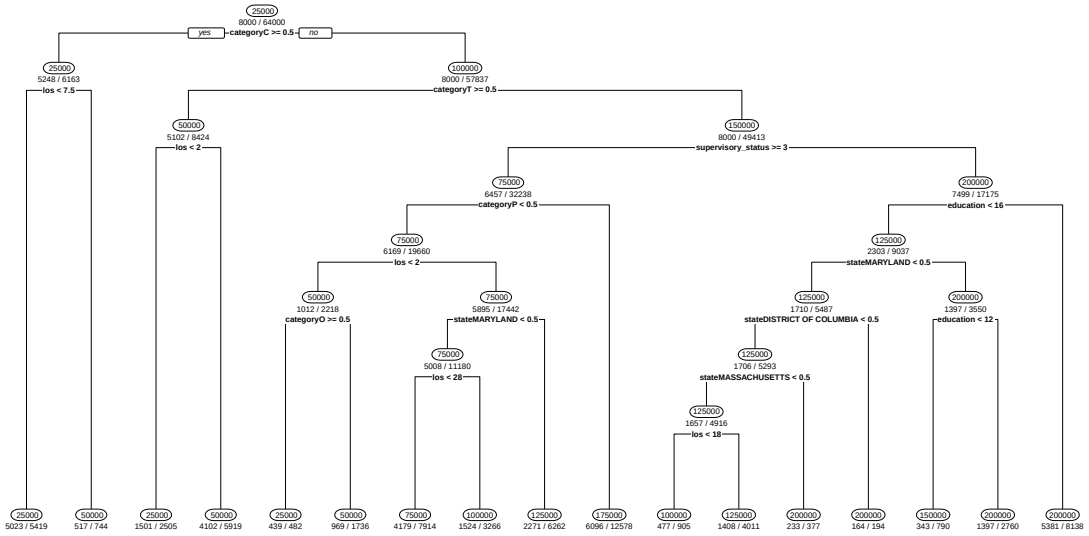

Figure 9: Pay Prediction SSA CART Model

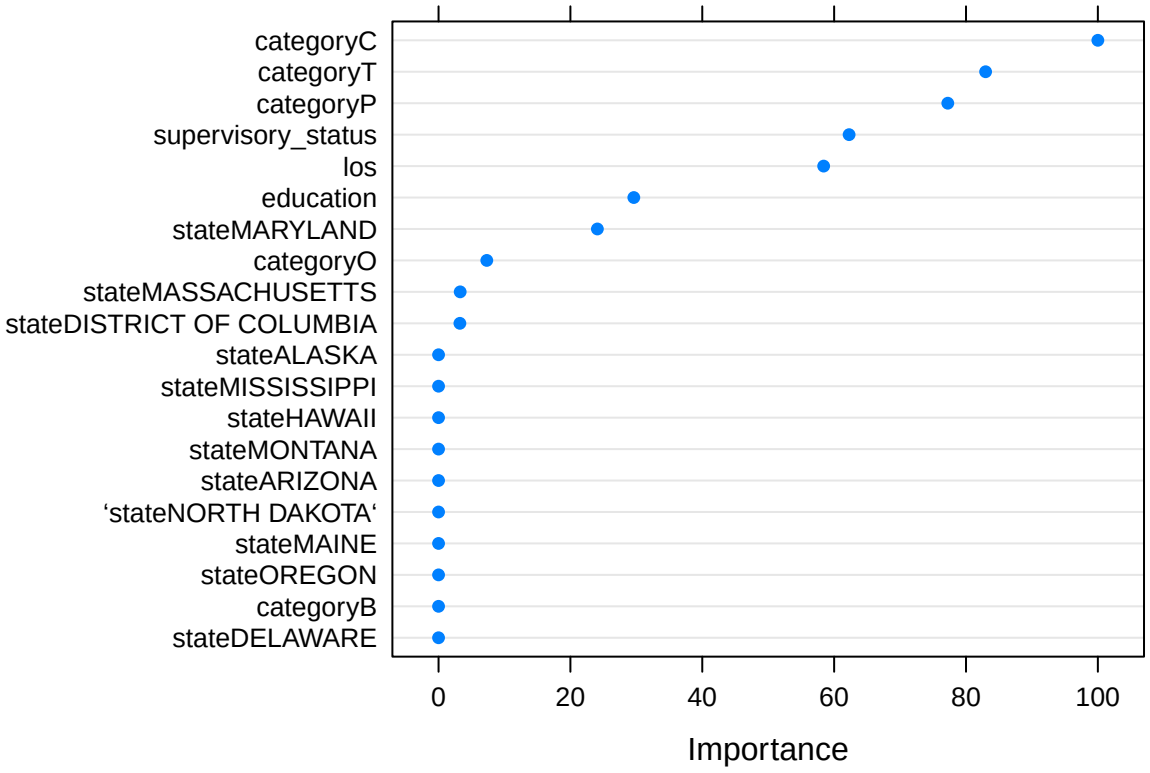

Figure 10: Pay Prediction SSA Variable Importance

Figures 9 and 10 show that the SSA model follows the same trend as the IRS model. However, the SSA has an additional category that is looks at and places more importance on supervisory status than length of service.

Table 18: SSA Pay Prediction Results

Accuracy	0.5668125
Kappa	0.5049286
AccuracyLower	0.5590917
AccuracyUpper	0.5745090
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	NaN

Table 18 shows that this model has an even higher accuracy score and kappa statistic than the IRS model.

Table 19: SSA Pay Prediction Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1780	357	2	0	1	0	0	0
50000	195	1370	349	52	22	7	2	0
75000	21	202	1072	550	108	23	11	31
100000	0	6	222	496	240	41	11	0
125000	4	61	274	651	953	507	127	99
150000	0	0	4	14	52	88	25	0
175000	0	4	64	144	376	912	1517	77
200000	0	0	13	93	248	422	307	1793

Table 20: SSA Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.8900	0.9742857	0.8317757	0.9841270	0.8317757	0.8900
Class: 50000	0.6850	0.9552143	0.6860290	0.9550096	0.6860290	0.6850
Class: 75000	0.5360	0.9324286	0.5312190	0.9336290	0.5312190	0.5360
Class: 100000	0.2480	0.9628571	0.4881890	0.8996263	0.4881890	0.2480
Class: 125000	0.4765	0.8769286	0.3561286	0.9214200	0.3561286	0.4765
Class: 150000	0.0440	0.9932143	0.4808743	0.8791174	0.4808743	0.0440
Class: 175000	0.7585	0.8873571	0.4903038	0.9625755	0.4903038	0.7585
Class: 200000	0.8965	0.9226429	0.6234353	0.9842274	0.6234353	0.8965

Tables 19 and 20 show that the predictions are more precise, reenforcing the idea that focusing on specific agencies in prediction is more useful than looking at the data as a whole.

Veteran's Health Administration

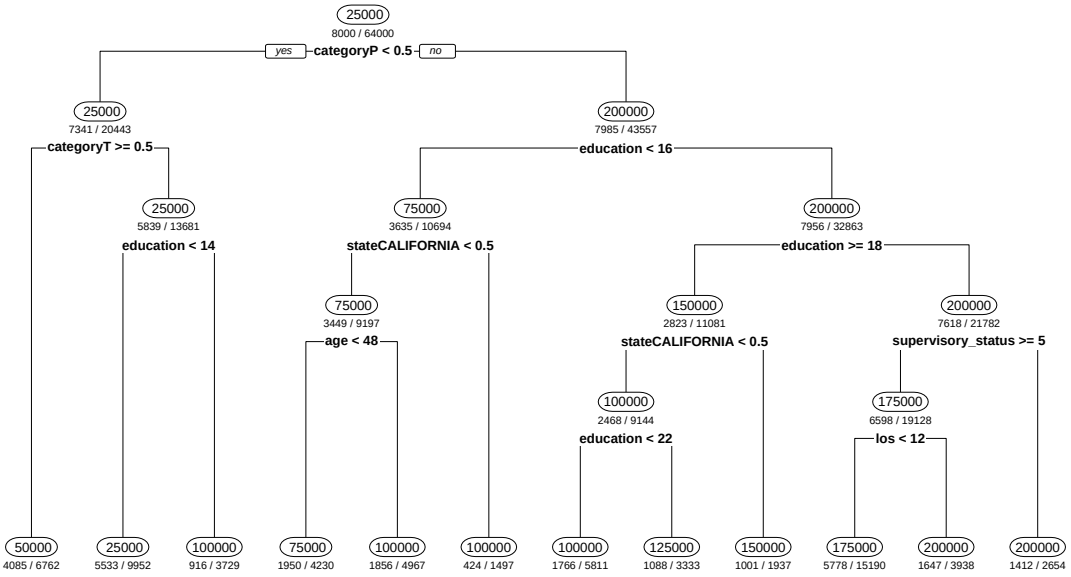

Figure 11: Pay Prediction VHA CART Model

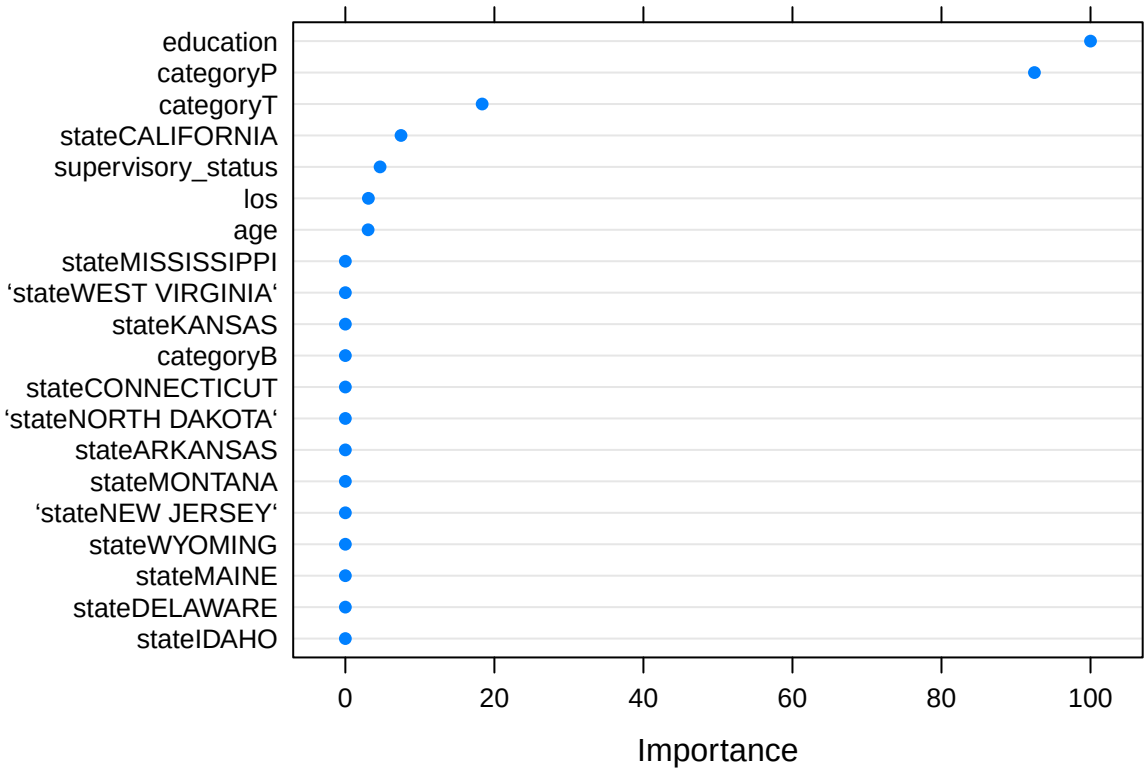

Figure 12: Pay Prediction VHA Variable Importance

Figures 11 and 12 show that education is the most important factor in the VHA. This is different from the IRS and SSA models which didn't seem to care much about education. Category is still an important factor in this model but surprisingly, so is the state of California. This does not make sense considering

what we know about how location should not affect pay in the federal government.

Table 21: VHA Pay Prediction Results

Accuracy	0.4228125
Kappa	0.3403571
AccuracyLower	0.4151413
AccuracyUpper	0.4305118
AccuracyNull	0.1250000
AccuracyPValue	0.0000000
McnemarPValue	NaN

Table 21 shows us that the accuracy score and kappa statistic of this model are similar to those of the general pay prediction model. This could explain why the state of California appeared as an important factor in this model. The other features were varied enough that a feature that should not have been important at all became important in comparison.

Table 22: VHA Pay Prediction Confusion Matrix

	25000	50000	75000	100000	125000	150000	175000	200000
25000	1387	717	241	86	35	17	4	0
50000	354	986	248	31	5	7	0	0
75000	73	97	488	264	77	36	5	6
100000	179	161	943	1208	701	683	136	19
125000	1	6	28	205	252	220	82	66
150000	4	1	20	73	93	258	35	13
175000	2	31	29	119	538	520	1430	1140
200000	0	1	3	14	299	259	308	756

Table 23: VHA Pay Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 25000	0.6935	0.9214286	0.5577000	0.9546363	0.5577000	0.6935
Class: 50000	0.4930	0.9539286	0.6045371	0.9294314	0.6045371	0.4930
Class: 75000	0.2440	0.9601429	0.4665392	0.8988899	0.4665392	0.2440
Class: 100000	0.6040	0.7984286	0.2997519	0.9338346	0.2997519	0.6040
Class: 125000	0.1260	0.9565714	0.2930233	0.8845443	0.2930233	0.1260
Class: 150000	0.1290	0.9829286	0.5191147	0.8876347	0.5191147	0.1290
Class: 175000	0.7150	0.8300714	0.3754266	0.9532442	0.3754266	0.7150
Class: 200000	0.3780	0.9368571	0.4609756	0.9133705	0.4609756	0.3780

Tables 22 and 23 show that this model does not have as many extremely inaccurate predictions as the general model but it does have a lot more than the IRS and SSA models. This tells us that looking at specific agencies rather than the entire dataset is only useful in certain cases where the data in the agency does not vary too much.

Education Prediction

In addition to predicting pay, I wanted to see if this data would be useful in predicting education. The CART method is used in this case as well because it is flexible in handling different types of data and is great for visualizing the detailed relationships between features. Since there are so many different education classes, I balanced the data by downsampling 5,000 instances of each class and then splitting the data into training and testing sets.

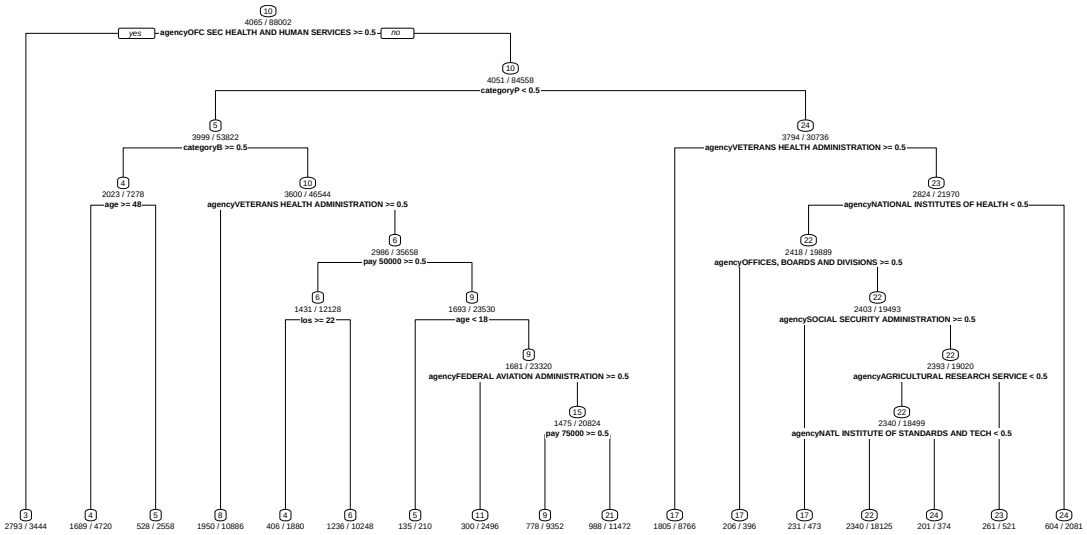

Figure 13: Education Prediction CART Model

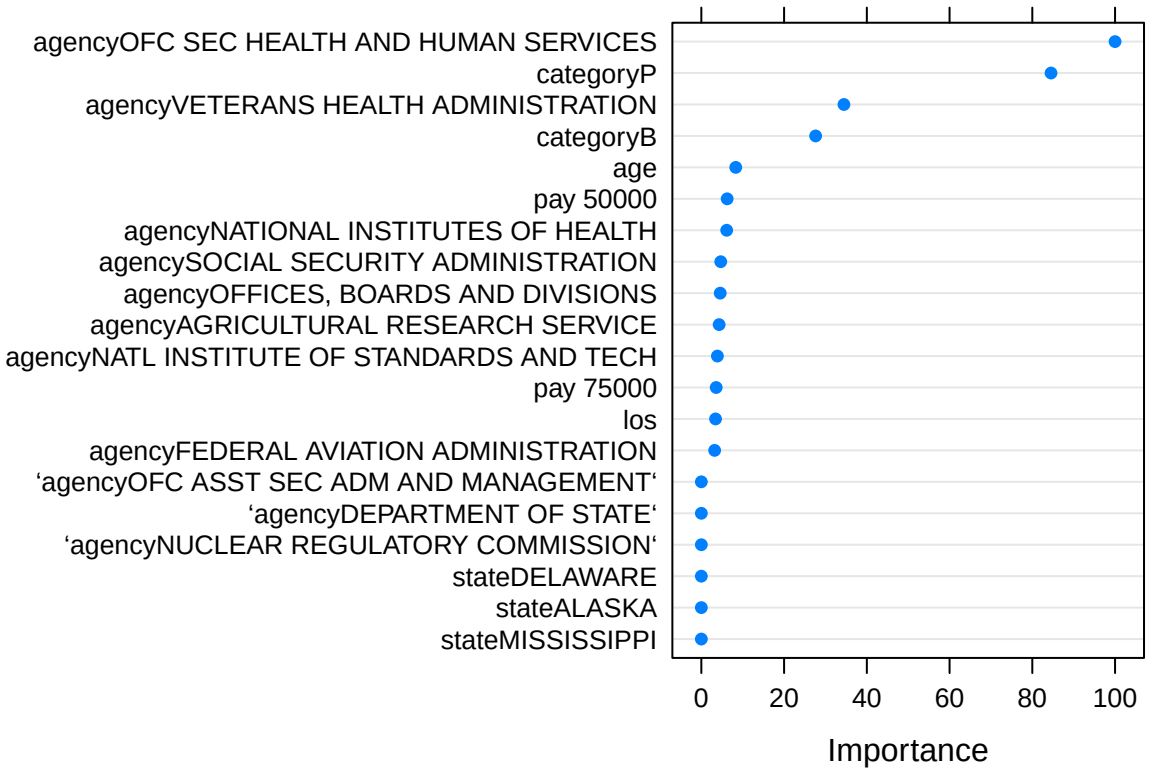

Figure 14: Education Prediction Variable Importance

Figures 13 and 14 show that agency and category are the most important features of this model. This might be because pay and length of service vary a lot even within educational levels so the model does not know what to do with them. On the other hand, some types of work are only available to people that have a certain education level.

Table 24: Education Prediction Results

Accuracy	0.1911992
Kappa	0.1520688
AccuracyLower	0.1860209
AccuracyUpper	0.1964592
AccuracyNull	0.0504137
AccuracyPValue	0.0000000
McnemarPValue	NaN

Table 25: Education Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 3	0.7244995	0.9921715	0.8225108	0.9862864	0.8225108	0.7244995
Class: 4	0.5070707	0.9453542	0.3042424	0.9760173	0.3042424	0.5070707
Class: 5	0.1595855	0.9725194	0.2103825	0.9618640	0.2103825	0.1595855
Class: 6	0.3320236	0.8959962	0.1341270	0.9650888	0.1341270	0.3320236
Class: 7	0.0000000	1.0000000	NaN	0.9550414	NA	0.0000000
Class: 8	0.4769833	0.8943128	0.1801036	0.9723230	0.1801036	0.4769833
Class: 9	0.1916996	0.8991232	0.0839463	0.9584497	0.0839463	0.1916996
Class: 10	0.0000000	1.0000000	NaN	0.9574961	NA	0.0000000
Class: 11	0.0663943	0.9744041	0.1077944	0.9572797	0.1077944	0.0663943
Class: 12	0.0000000	1.0000000	NaN	0.9544959	NA	0.0000000
Class: 13	0.0000000	1.0000000	NaN	0.9546777	NA	0.0000000
Class: 14	0.0000000	1.0000000	NaN	0.9495863	NA	0.0000000
Class: 15	0.0000000	1.0000000	NaN	0.9556778	NA	0.0000000
Class: 16	0.0000000	1.0000000	NaN	0.9526321	NA	0.0000000
Class: 17	0.5773399	0.9159319	0.2493617	0.9781657	0.2493617	0.5773399
Class: 18	0.0000000	1.0000000	NaN	0.9562233	NA	0.0000000
Class: 19	0.0000000	1.0000000	NaN	0.9562233	NA	0.0000000
Class: 20	0.0000000	1.0000000	NaN	0.9555414	NA	0.0000000
Class: 21	0.2381871	0.8740518	0.0855559	0.9586626	0.0855559	0.2381871
Class: 22	0.5937193	0.8095238	0.1314932	0.9762028	0.1314932	0.5937193
Class: 23	0.0656250	0.9970054	0.5000000	0.9589887	0.5000000	0.0656250
Class: 24	0.2089704	0.9816815	0.3474576	0.9637519	0.3474576	0.2089704

Tables 24 and 25 show that this model performs terribly and many classes are not predicted at all. Even 13 is never predicted which, as seen in Project 1, is one of the most frequently occuring education classes. This model would not be useful in any case since it seems to be prone to guessing either really low education levels or really high even though the highest occurences would be in the middle.

Supervisory Status Predicion

I also wanted to see if this data would be useful in predicting supervisory status. The CART method is used and I balanced the data by downsampling 20,000 instances of each class and then splitting the data into training and testing sets. I did this because there are less possible values for supervisory status or pay and education. Using a decision tree for this model is a good idea because of the flexibility of how it handles different data types as well as how its simplicity and quick training time are great for models that don't require high precision.

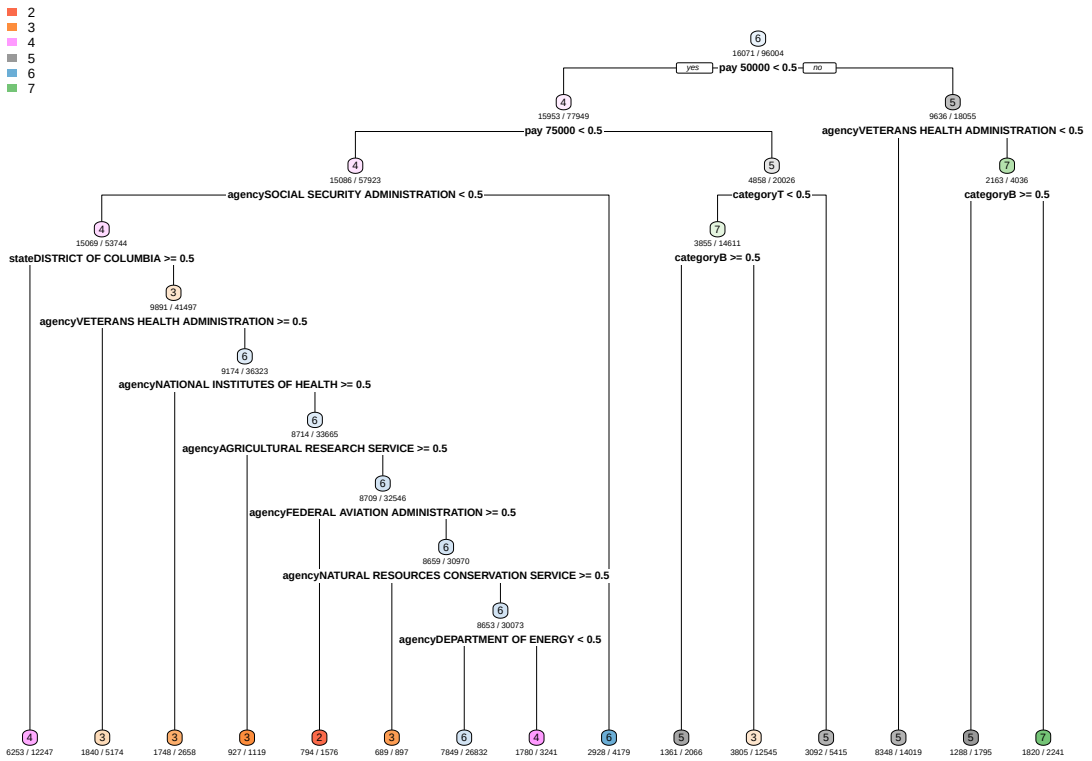

Figure 15: Supervisory Status Prediction CART Model

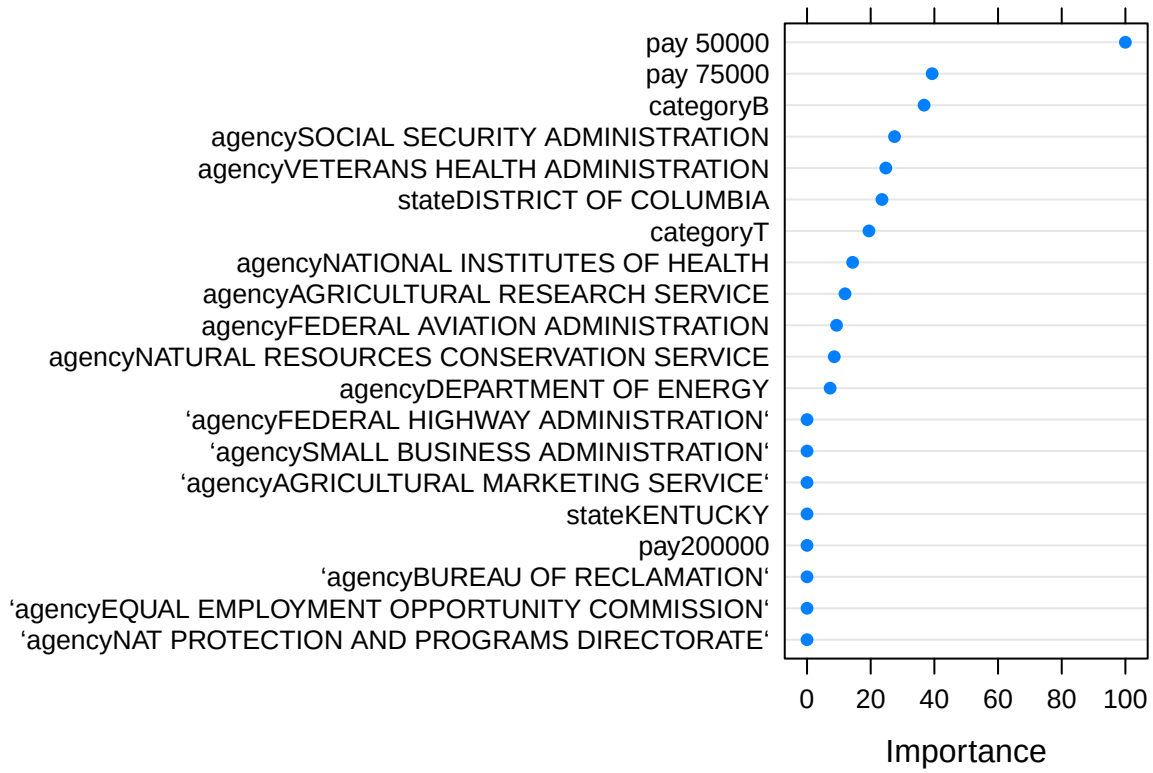

Figure 16: Supervisory Status Prediction Variable Importance

Figures 15 and 16 show that pay and agency are the best indicators of supervisory status. This makes sense because higher levels of authority almost always go along with a higher salary. This also indicates that specific agencies have higher levels of authority in general than others.

Table 26: Supervisory Status Prediction Results

Accuracy	0.4642440
Kappa	0.3576269
AccuracyLower	0.4579178
AccuracyUpper	0.4705790
AccuracyNull	0.1681114
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

Table 27: Supervisory Status Prediction Confusion Matrix

	2	3	4	5	6	7
2	212	17	23	3	9	141
3	1088	2273	333	150	477	1311
4	571	221	1998	132	616	245
5	602	337	3	3559	161	1191
6	1476	1109	1675	149	2666	714
7	52	50	0	0	0	432

Table 28: Supervisory Status Prediction Class Statistics

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 2	0.0529868	0.9903476	0.5234568	0.8393879	0.5234568	0.0529868
Class: 3	0.5672573	0.8319576	0.4035866	0.9055761	0.4035866	0.5672573
Class: 4	0.4955357	0.9105891	0.5281523	0.8993717	0.5281523	0.4955357
Class: 5	0.8913098	0.8853172	0.6080642	0.9760789	0.6080642	0.8913098
Class: 6	0.6785442	0.7447052	0.3422776	0.9220707	0.3422776	0.6785442
Class: 7	0.1070897	0.9948903	0.8089888	0.8464752	0.8089888	0.1070897

Tables 26, 27, and 29 show that the performance of the supervisory status model are similar to the general pay prediction model. This indicates the close relationship between the two features. However, this model is not very useful since an employee cannot decide their own supervisory status.

Evaluation and Deployment

The models created here have fairly low accuracies but when taking into account the fact that false predictions are typically fairly close to the actual values, these models could still be useful for prospective government employees. They can get a fair estimate of how much they can expect to receive in compensation in certain departments and states based on their qualifications.

These models are also extremely transparent and show how each factor affects pay. This helps employees make detailed decision about where to work or whether or not to go back to school based on their specific circumstances. We saw in figure 2 that employees with a lower education level could expect to make a lot more money by working for the Federal Aviation Administration.

Current employees can also asses these models to see if they are being paid a fair salary in comparison to others with the same qualifications and in the same circumstances. They can also see if they make similar salaries to those of the same authority levels.