Predicting Job Mismatch Rate Through Job Satisfaction - PPOL 670 Group Project

Connor Harrison, Dong Hoon Lee, Haorui Sun

April 28, 2019

Method Section

The first approach we took in answering our hypothesis is using data visualizations to look for patterns and relationships between the dependent variable and independent variable while controlling for difference in characteristics such as region, race, highest degree attained, job sector, age, and company size. And because survey data set contained one variable for overall satisfaction at the job and one for each aspects of the job, we decided to use all the job satisfaction variables for the visualization analysis to look for additional insights about different types of job satisfaction.

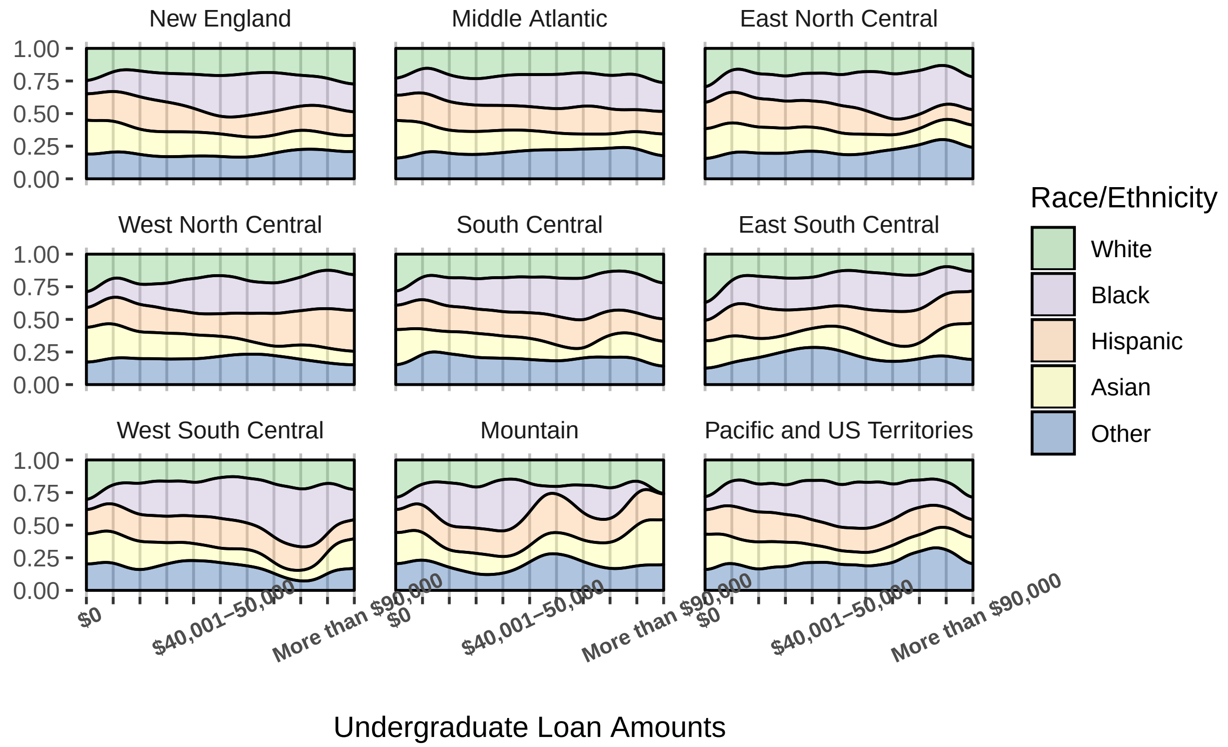
The first step to visualization analysis was to recode variables to be suitable for visualization. The original satisfaction variables had the following four possible responses: "very dissatisfied," "dissatisfied," "satisfied," and "very satisfied." For the purposes easier presentation and interpretation, we created an indicator for those who reported as "satisfied" or "very satisfied" with their job or aspects of the job. The key independent variable, undergraduate student loan amount, was also ordinally categorized with the starting loan amount of $0 and going up in the increments of $10,000 up to $90,000, at which point was grouped in to more than $90,000. Because the values of this variable were not numerical to begin with, we grouped loan amounts into increments of $30,000 (e.g. between $1 - $30,000 and $30,000 - 60,000) with the last value group consisting of those with more than $90,000 of loan.

We first looked at the overall distribution of loan amounts by race groups and region to see if there's any significant differences in terms of loan borrowing trends. Then we plotted overall job satisfaction level by loan amount groups in different regions and then by race and ethnicity groups. Lastly, we plotted the average satisfaction level in their jobs and in various aspects of it in the y-axis and undergraduate loan amounts in the x-axis, while faceting them over covariates such as region, race, and highest degree.

Discussion Section

In Figure 1, we observed that white college graduates in most regions have the highest proportion of zero loan holders among all the race groups and their share gets smaller with increasing loan amounts. On the other hand, black and Hispanic graduates' share of loan holders seems to increase with loan amount, suggesting that there exists a debt disparity among races. Particularly, blacks in West North Central, South Central and West South Central regions showed large increase in their share as their loan amount increased.

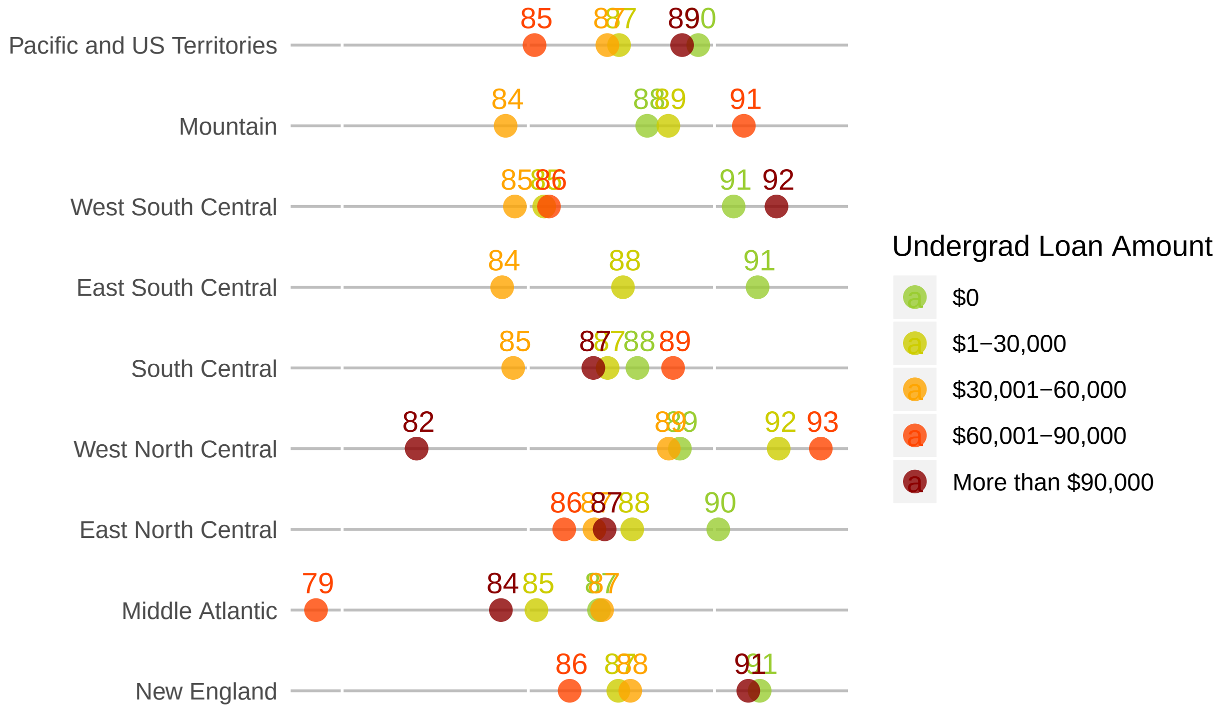
Figure 1. Undergraud Loan Borrowing Trend by Employer Region and Respo



Data: National Survey of College Graduates

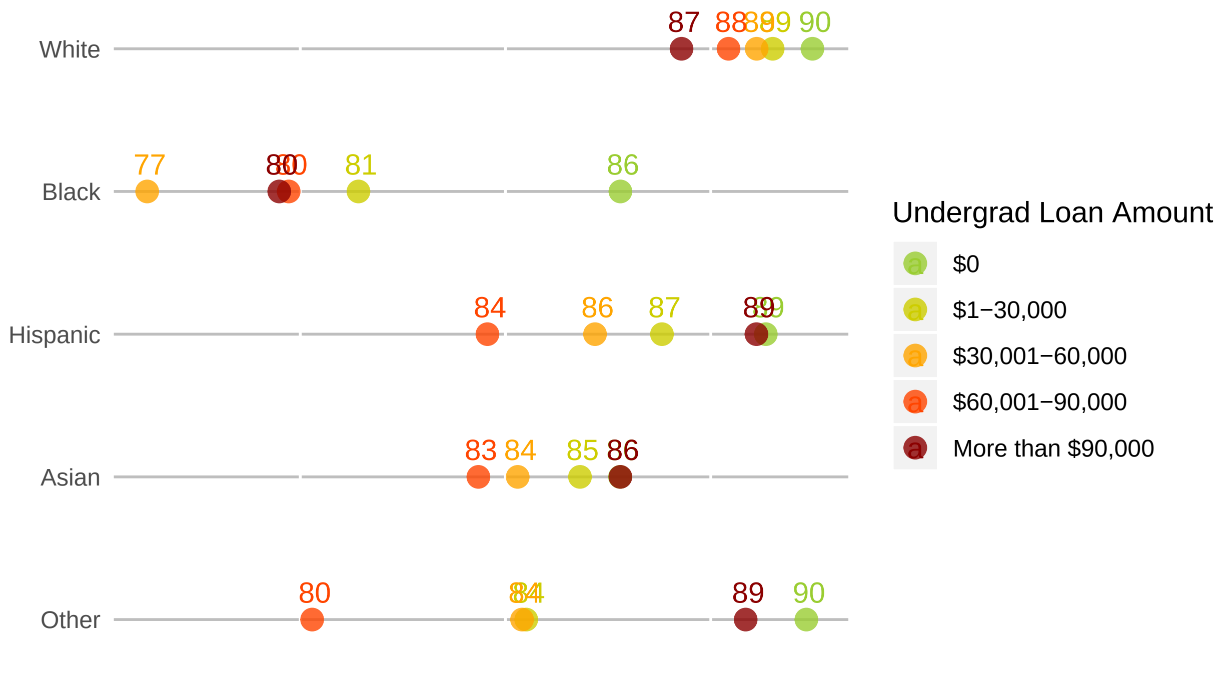
In Figure 2, overall job satisfaction was generally lower for those with higher loan amounts than those with any loans. However, there were few instances where job satisfaction was highest for the higher loan amount groups such as for those working in the Mountain, West South Central, South Central, and West North Central regions. Figure 3 also revealed an interesting pattern of the highest loan amount group ($90,000 or more) having almost as high satisfaction levels as those with zero loans for Hispanics, Asians, and other race groups. This trend of higher satisfaction levels for those with very high loan amounts suggests a quadratic relationship where the satisfaction level drops when you go from no loans to some loans and goes up again after certain amounts of loan. Another interesting observation we made was that variation in satisfaction levels are much wider in certain groups. For instance, when we look at by race, the variation in the average satisfaction level by loan amounts is very small for white and Asian respondents with data range of about 3 percentage points, whereas black respondents had a range of about 9 percentage points and other race respondents with 10 percentage points. This indicates that loan amounts many affect job satisfaction differently by race or factors correlated with race.

Figure 2. Percent Satisfied with Job by Undergraduate Loan A



Data: National Survey of College Graduates

Figure 3. Percent Satisfied with Job by Undergraduate Loan Amount and I



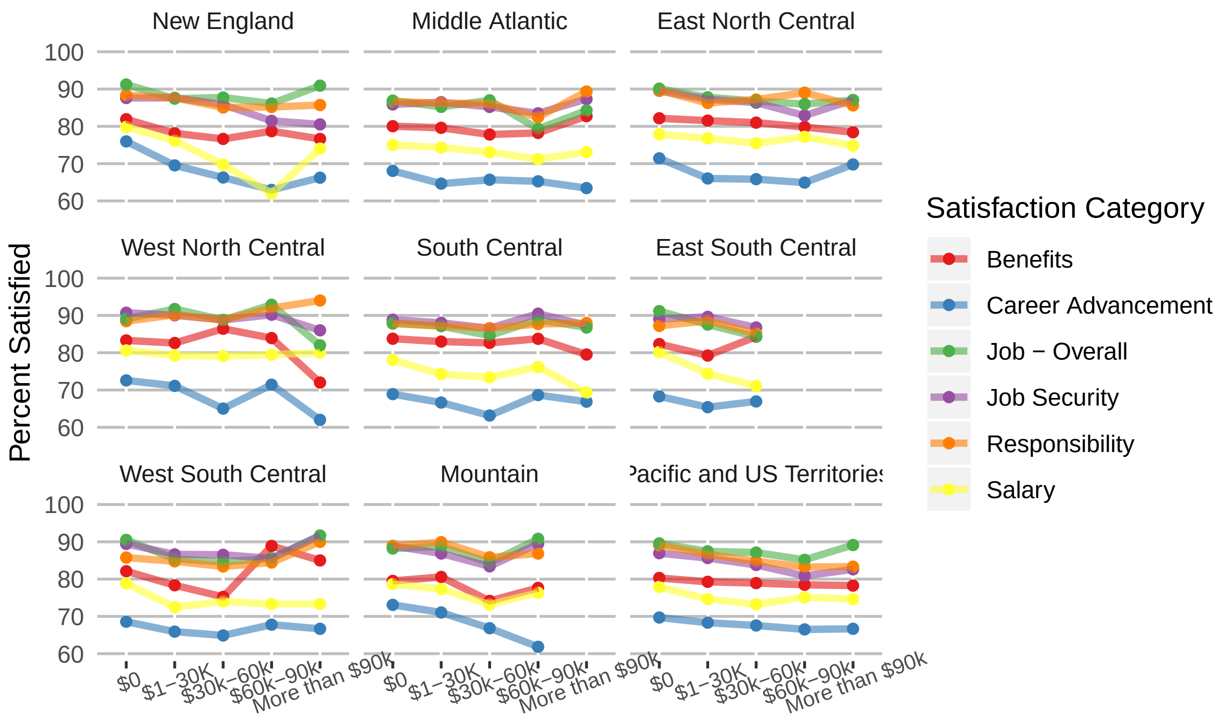
Data: National Survey of College Graduates

Figure 4 through 8 are showing the changes in satisfaction levels over loan amounts by various characteristics of the respondents. Overall, these visualizations did not show a singular relationship between undergraduate loan amounts and satisfaction levels. Some analysis groups showed almost no change in satisfaction levels with increasing loan amounts, whereas others showed random spikes or dips in the satisfaction levels. However, a quadratic relationship that was also noted in Figure 2 was also observed in several of the analysis groups, suggesting that a non-linear relationship between loan amounts and satisfaction levels.

Satisfactions in the job, security, and responsibility were the highest among all job satisfaction categories and they aligned with each other in terms of satisfaction levels and degree of change. Few of the plots, such as the plot for New England region, university jobs, and federal and state jobs, had one of the three satisfaction categories diverge off from each other. Regardless of these few exceptions, these figures seem to suggest that there is a high correlation among overall, security, and responsibility at the job.

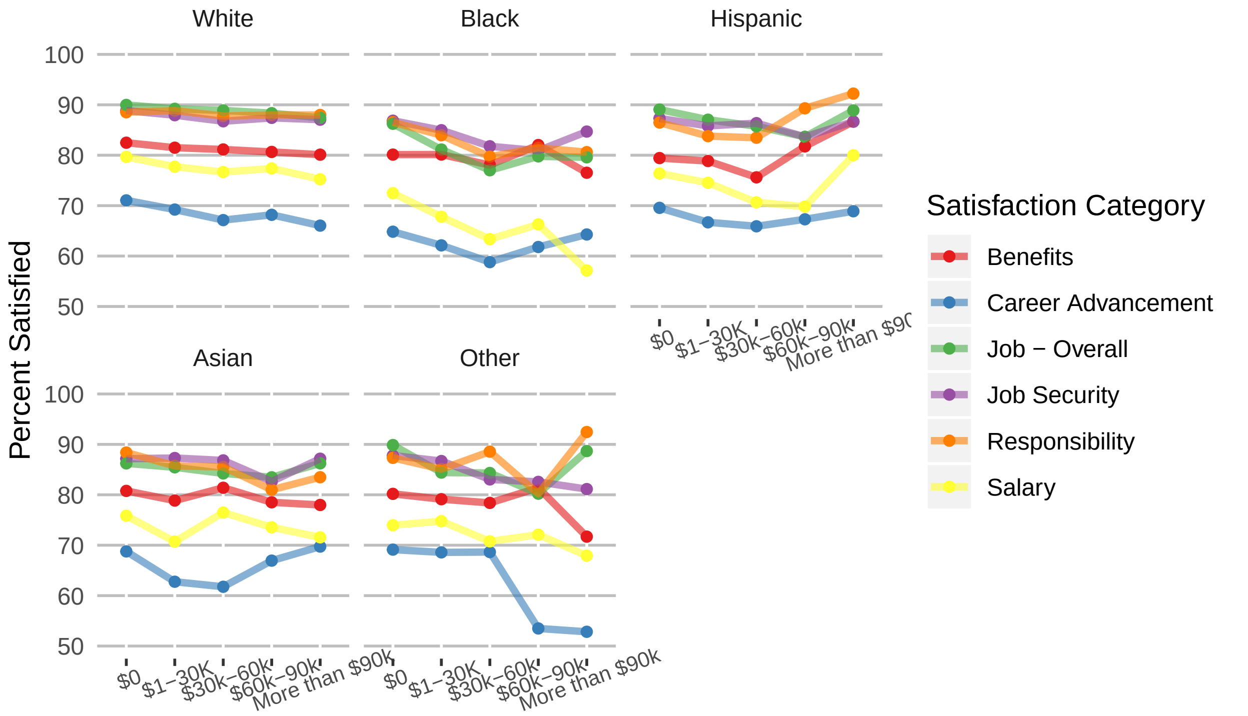
In some cases, we had satisfaction categories with generally lower levels to show higher satisfaction level than the overall satisfaction category as in the case of federal and state government jobs. In this group, the satisfaction levels for job benefits and security were noticeably higher than that of the overall satisfaction for all undergraduate loan amounts. That is an expected result considering that government jobs are well known for their job security and benefits.

Figure 4. Types of Job Satisfaction by Undergraduate Loan Amounts and b



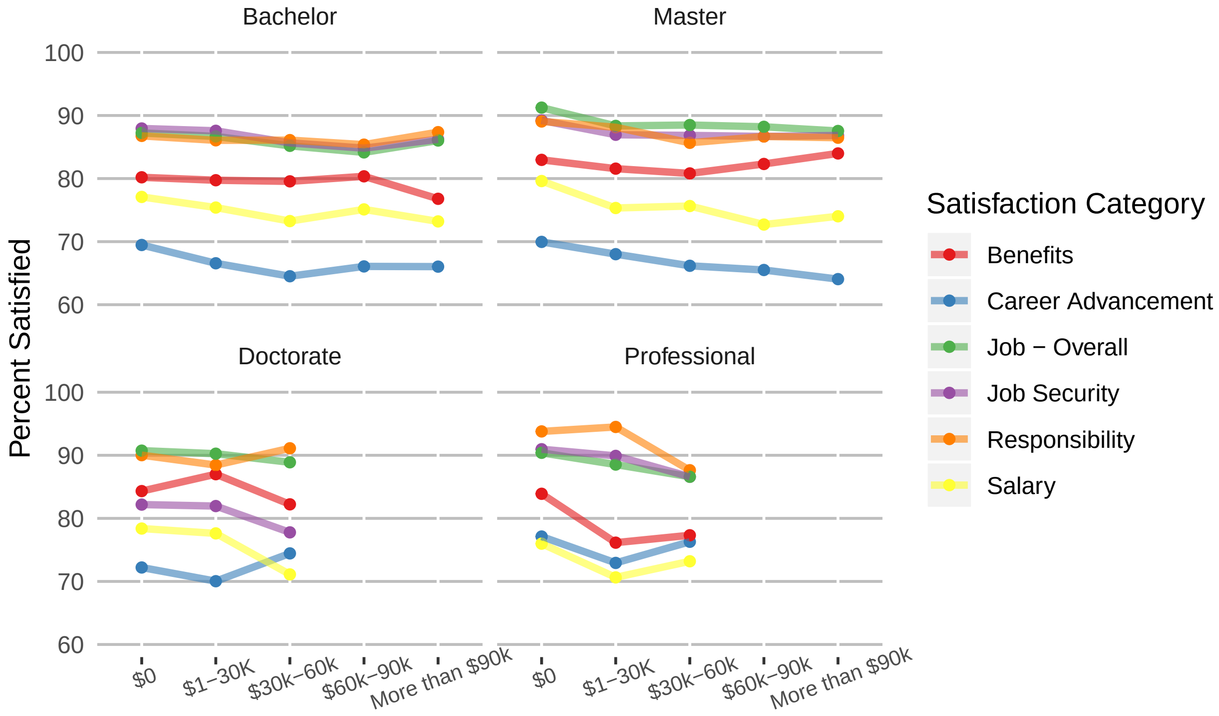
Data: National Survey of College Graduates

Figure 5. Types of Job Satisfaction by Undergraduate Loan Amounts and b



Data: National Survey of College Graduates

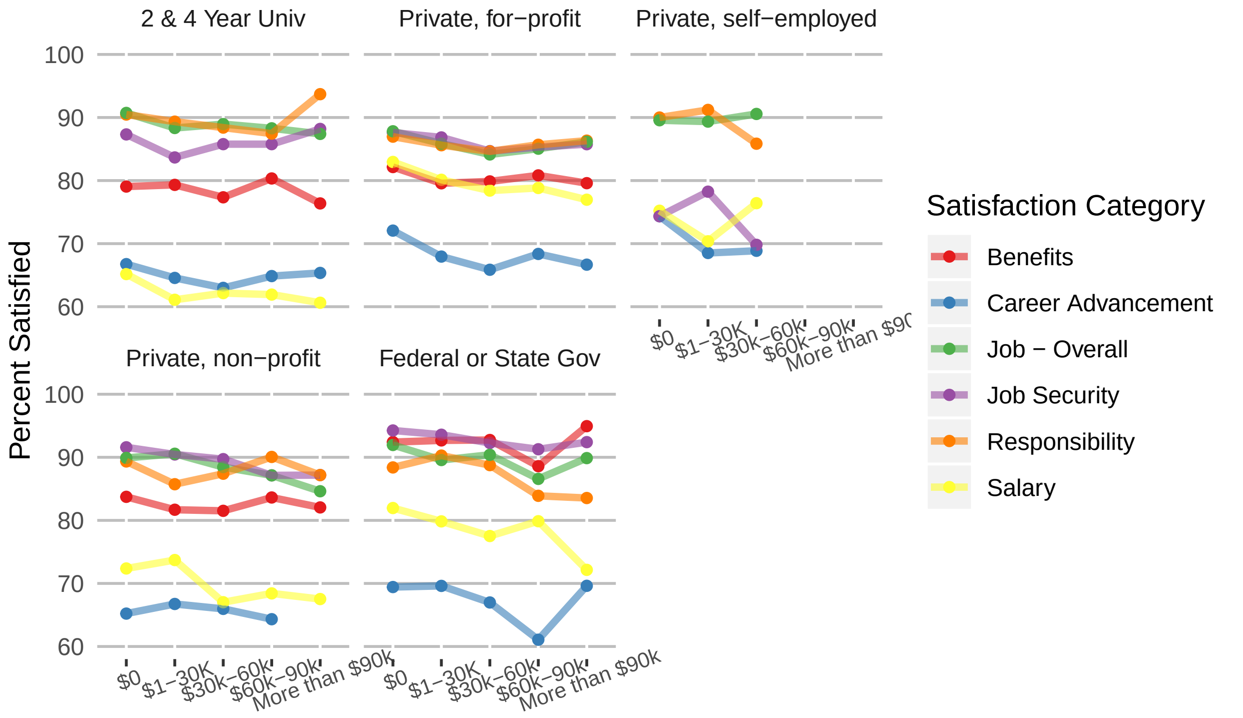
Figure 6. Types of Job Satisfaction by Undergraduate Loan Amounts and H



Data: National Survey of College Graduates

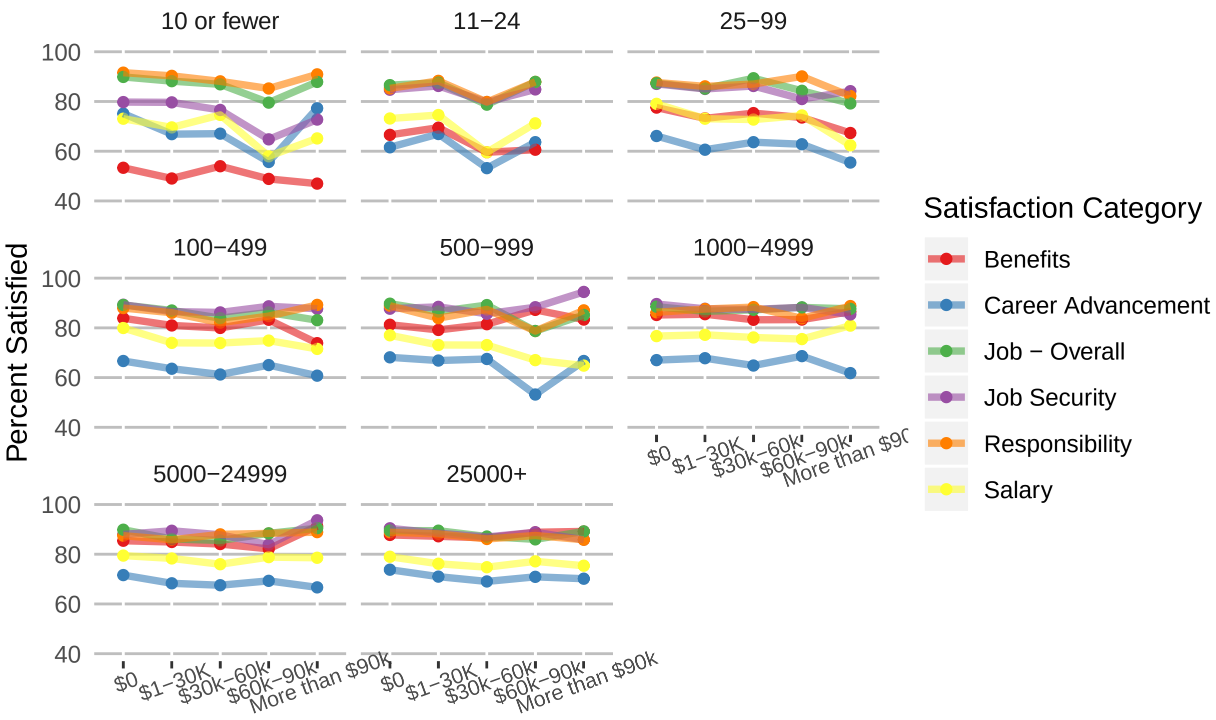
```
## Warning: Removed 4 rows containing missing values (geom_point).
```

Figure 7. Types of Job Satisfaction by Undergraduate Loan Amounts and b



Data: National Survey of College Graduates

Figure 8. Types of Job Satisfaction by Undergraduate Loan Amounts and b

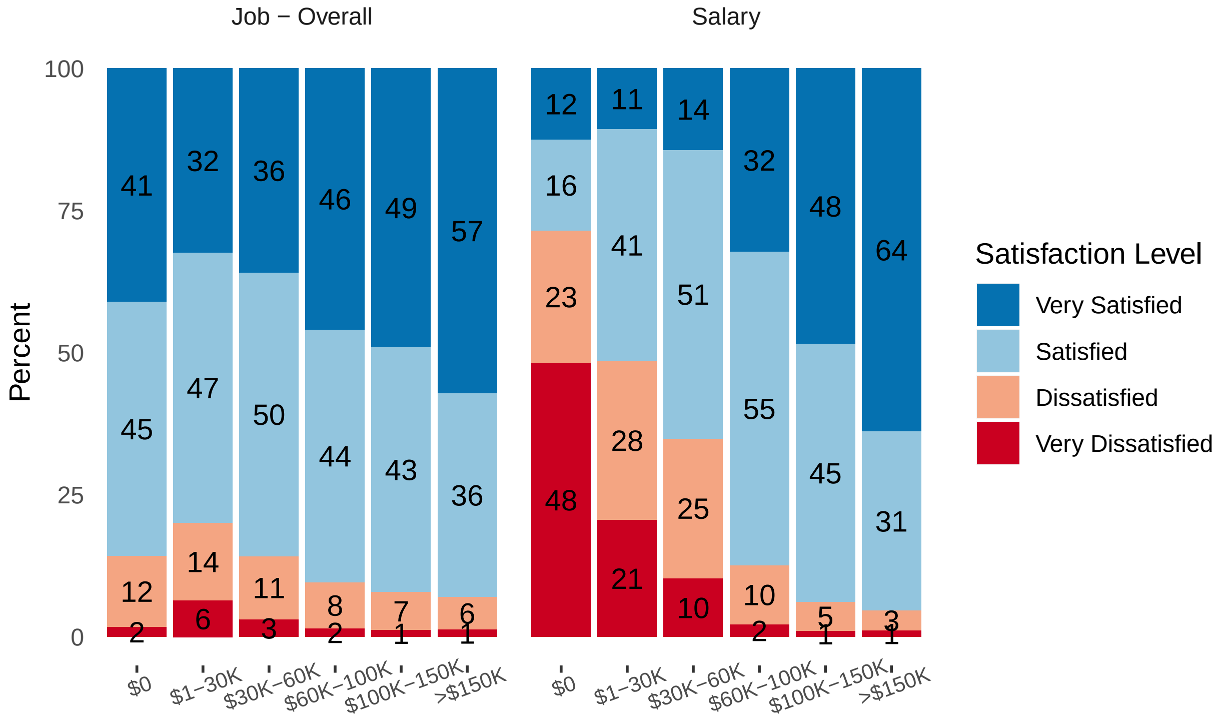


Data: National Survey of College Graduates

In figures 9 and 10, we created bar charts to show the distribution of the responses for job satisfaction questionnaire over salary and age groups. The percent of satisfied and very satisfied responses increased with increasing salary amounts and age. However, the changes were very small and, therefore, difficult to be claimed as meaningful.

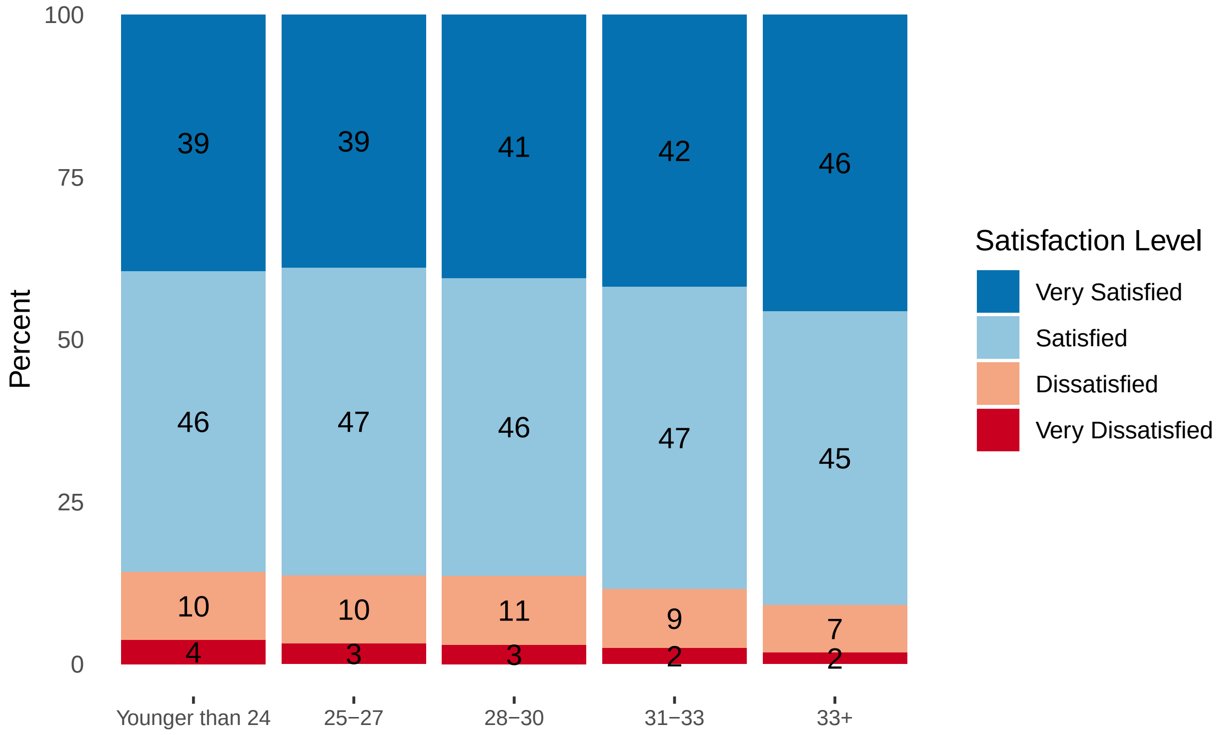
Through this visual analysis method, we found few patterns that suggested quadratic relationships between our dependent and independent variables. However, the satisfaction levels were generally very high (around 85%) and had small variations in them, which meant that there was little room for meaningful variations to be observed. And because this analysis is limited to three variables at a time, these relationships cannot be interpreted as casual or definitive as there could be omitted variable bias that is affecting the relationships observed.

Figure 9. Job and Salary Satisfaction by Salary Groups



Data: National Survey of College Graduates

Figure 10. Job Satisfaction by Age Groups

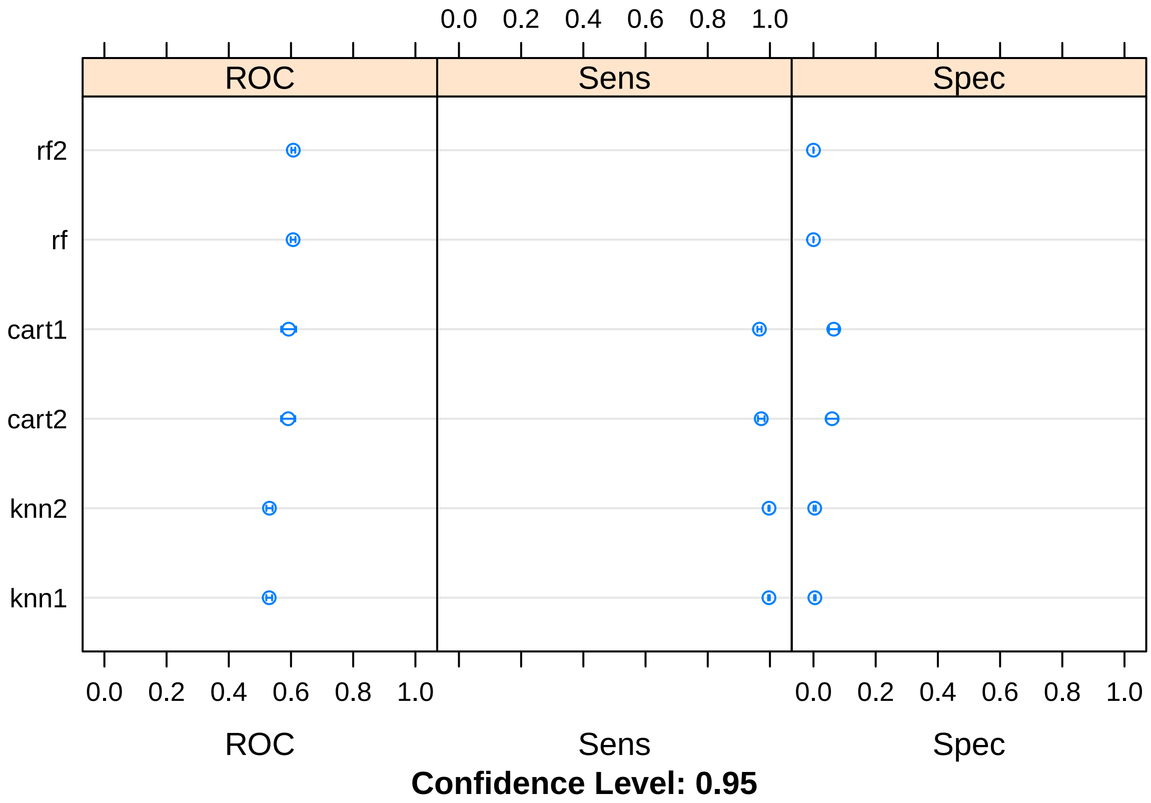


Data: National Survey of College Graduates

# Using Machine Learning to Identify Predictors of Job and Salary Satisfaction Now that we've identified the primary variables that can potentially influence job and salary satisfaction, we can use machine learning to predict which outcomes and characteristics of recent graduates influence these outcomes.

Which Model predicts our outcome the best?

To compare which of the six machine learning models performed the best on the training data, we can create a dotplot, which creates an easy-to-assess visual of model performance across three measures: the area under the ROC curve (or total predictive power), sensitivity, and specificity. The plot below ranks the models by performance across these metrics.



Testing the Model

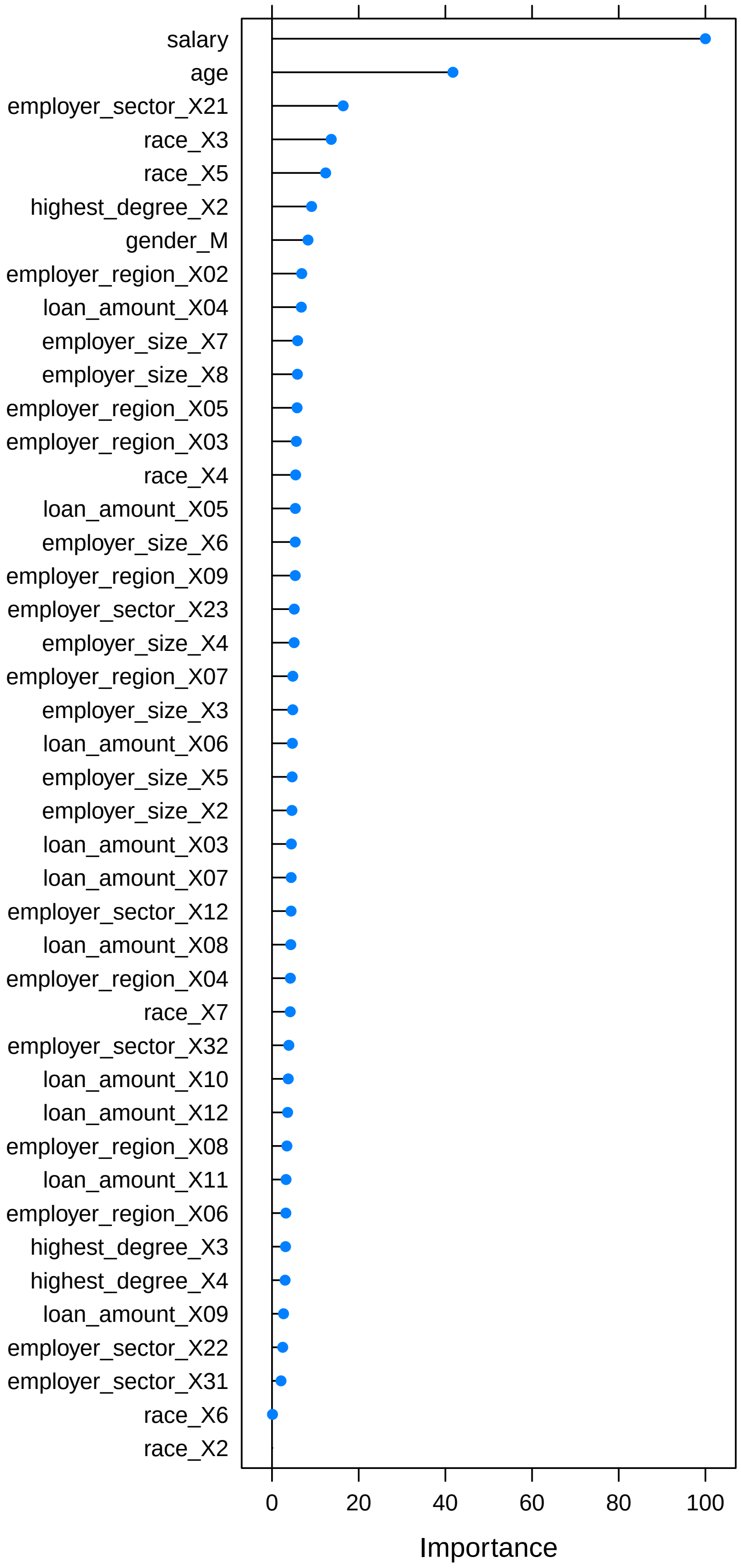
Now that we have run six different machine learning models - two iterations of each K Nearest Neighbor, Regression Tree, and Random Forest models - we can test the predictive accuracy of the model that performed the best on our training data, the tuned Random Forest model. To do so, we will use a "confusion matrix", which shows how accurate our model is in predicting true positives, false positives, false negatives, and true negatives. The results of the confusion matrix for the tuned Random Forest model are shown below:

```
## Confusion Matrix and Statistics
##
##
## pred             Satisfied Not_Satisfied
##   Satisfied           4851           680
##   Not_Satisfied          0             0
##
##                Accuracy : 0.8771
##                  95% CI : (0.8681, 0.8856)
##     No Information Rate : 0.8771
##     P-Value [Acc > NIR] : 0.5102
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
```

```
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.8771
##          Neg Pred Value :    NaN
##              Prevalence : 0.8771
##          Detection Rate : 0.8771
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Satisfied
##
```

Accuracy is the rate at which the model correctly identifies that an individual is satisfied with their job; specificity is the rate the model correctly classifies true negatives, and sensitivity, which is the rate at which the model correctly classifies all true positives. While this model is highly accurate with an accuracy rate of 87.7%, we can see there is a clear issue with the model's predictive ability as the sensitivity is 100% and specificity is 0%. The confusion matrix for this model shows that it predicted an individual is satisfied with their job 100% of the time, which means all true positives are identified and no true negatives were identified. The high accuracy of the model comes from the fact that individuals in the data were much more likely to report being satisfied with their job than unsatisfied. Therefore, because the 'positive' outcome is so prevalent, the model always predicts the positive outcome, making the accuracy of the model equal to the prevalence of the positive outcome in the data.

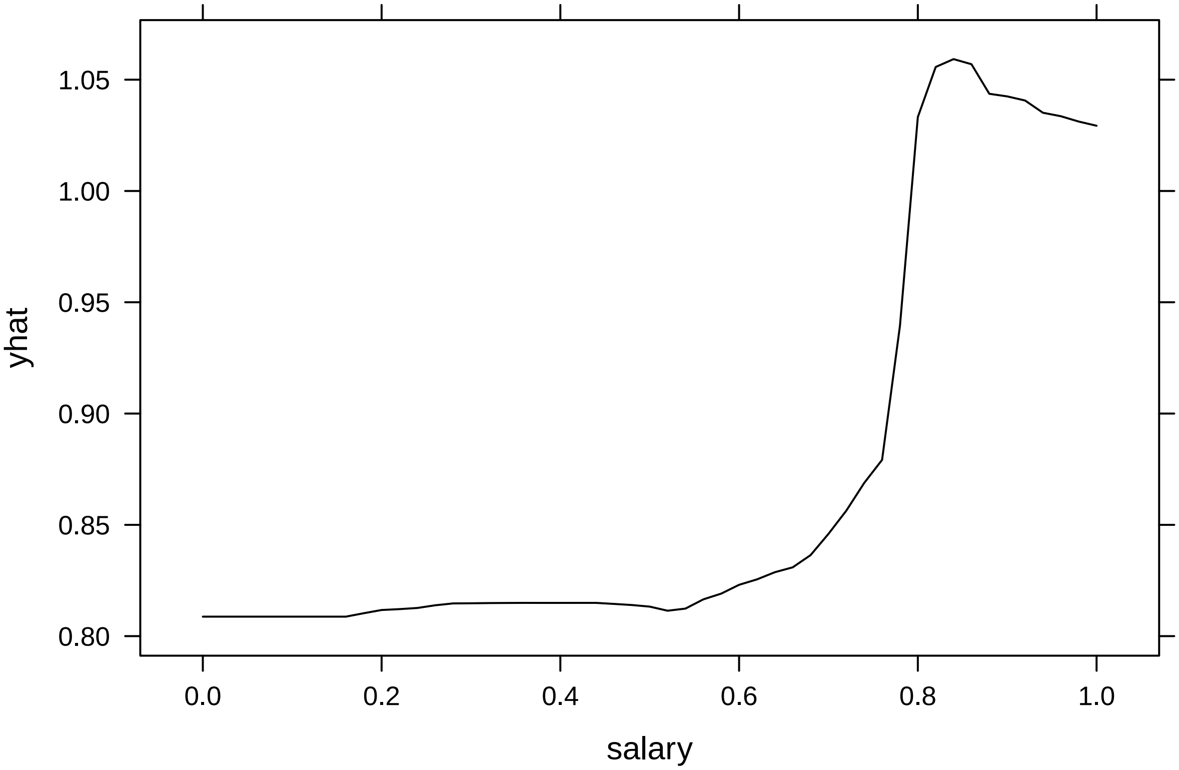
With this information in mind, we can now break down the model into individual predictive factors to examine which variables have the greatest influence over our outcome. In other words, what factors influence job satisfaction to the greatest degree? The following plot ranks each variable by predictive accuracy in measuring job satisfaction:



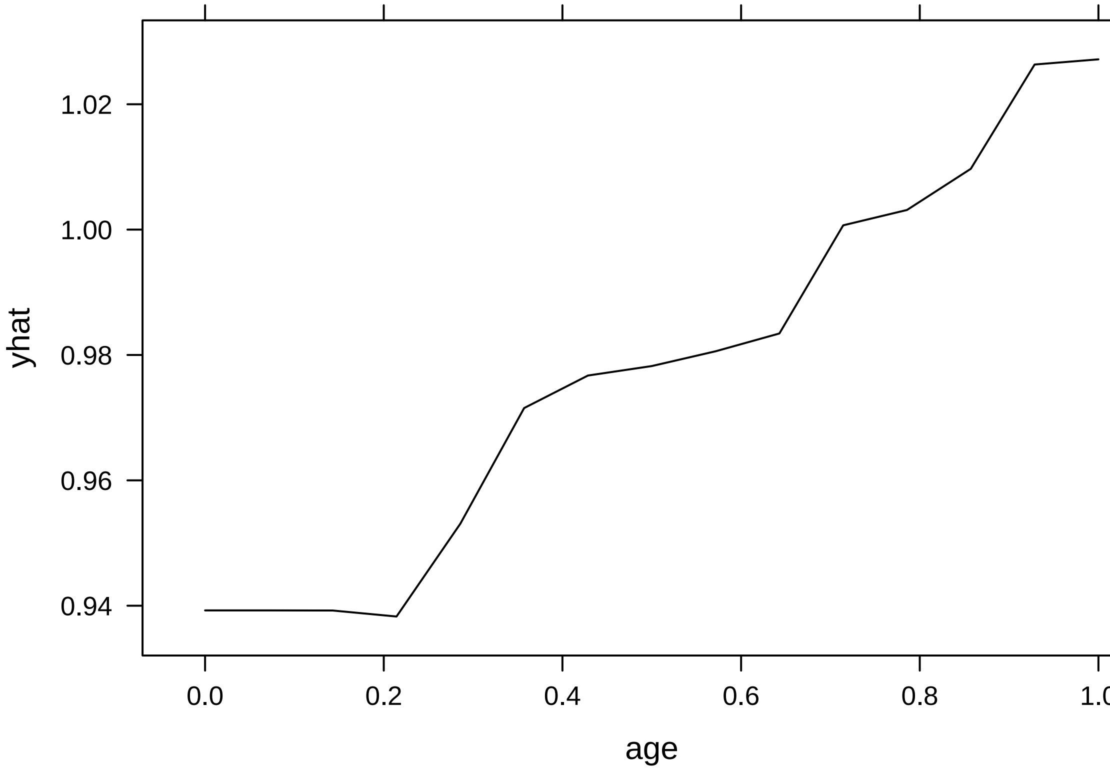
The variable importance plot

clearly shows that an individuals' salary perfectly predicts that individuals satisifaction with their job. While we expected salary to have an influence over an individual's job satisfaction, we did not expect it to have as large of an influence as it does. The second and third predictive factors are age and race (X5), respectively. This model reveals that undergraduate loan amount is not a strong predictor of job satisfaction, as the first of the loan indicators appears in the 9th position.

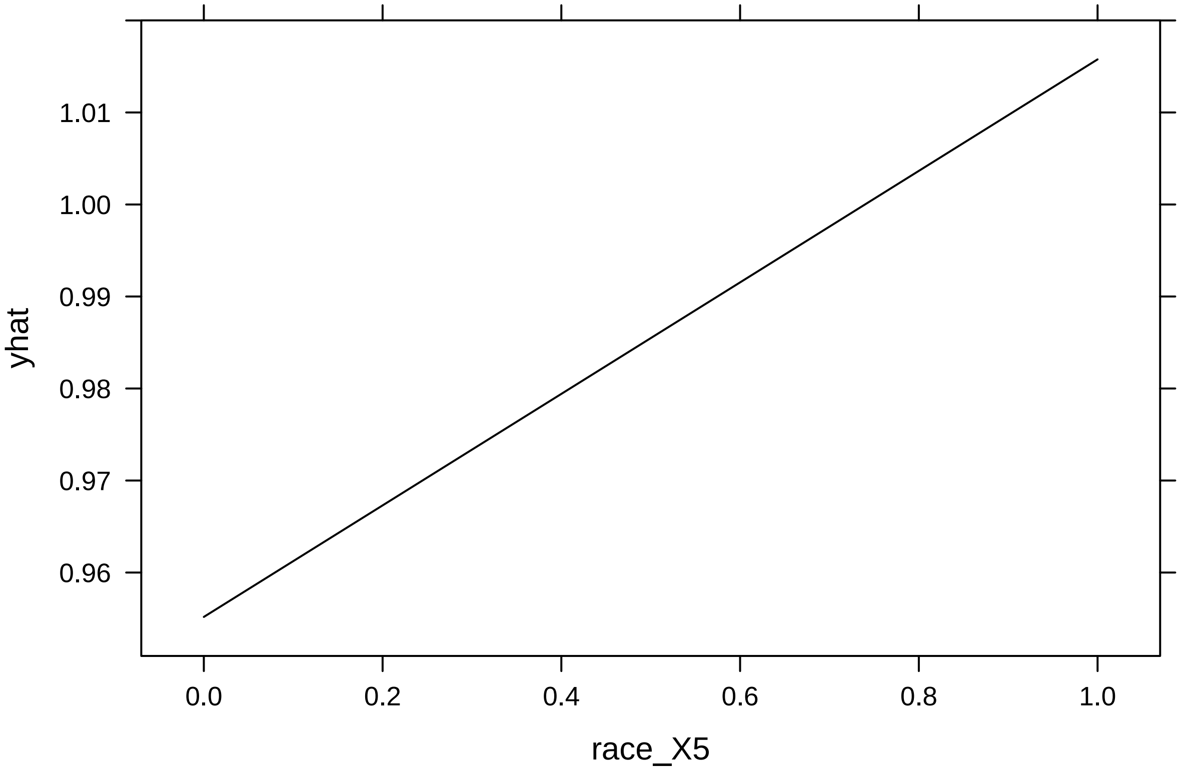
Having identified the variables that are the strongest predictor of job satisfaction in the best performing machine learning model, we can now examine the relationship between these variables and the outcome. To do so, we use partial dependency plots, which graph the predictive power of each variable at different levels of the variable and the directional relationship between the variables. Partial dependency plots for salary, age, and race (X5) are shown below:



Salary is a strong and constant predictor of job satisfaction until really high levels of salary, at which point the



predictive power falls away.

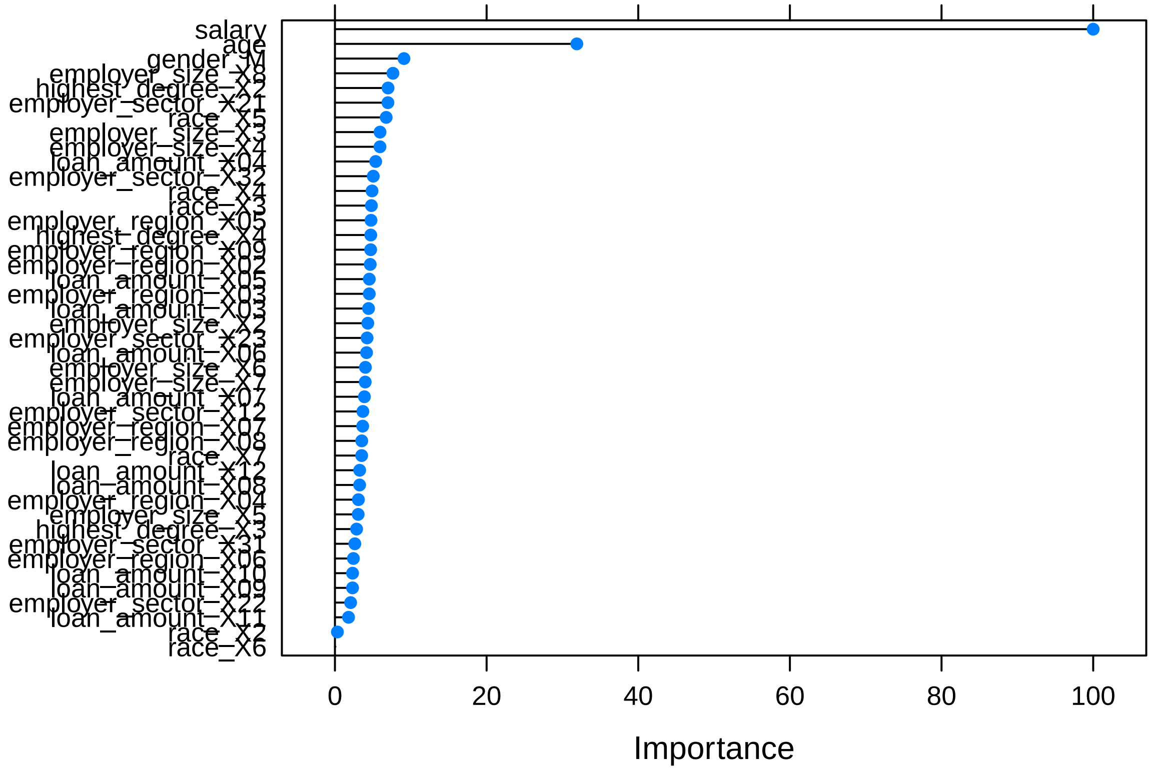


We checked the outcome of the random forest model by also testing the tuned K-Nearest Neighbor and tuned Regression Tree models. We wanted to see if perfect specificity was a quirk of the random forest model or if all the machine learning models leaned heavily on this predictive pattern. As it turns out, the output of the KNN and regression tree models are not significantly different than the RF model. The accuracy of both models is very similar to the RF model, yielding 87.6% and 87.7% accuracy, respectively.Similarly, both models had near perfect sensitivity and zero specificity.

Another way in which we tested this result was to increase the size of our sample. In the initial models and visualizations, we limited the sample to individuals younger than 36, as we predicted that recent trends in the cost of higher education were more likely to impact the millennial generation. Having discovered that the amount of loans a college graduate has does not help predict their level of job satisfaction, we determined it appropriate to expand the machine learning model to the entire sample. We did not find significantly different results using this larger sample. As it turns out, the percentage of individuals that report being satisfied with their job in the full sample is similar to the percentage in the original sample, and therefore the model acted in a similar fashion, with high accuracy, near perfect specificity, and almost no sensitivity.

Given that there appears to be very little variation in overall job satisfaction among respondents, we decided to examine other measures of job satisfaction. The survey also asked individuals how satisfied they are with specific aspects of their job. We chose to examine satisfaction with salary potential for career advancement. We examined salary satisfaction because of the importance of salary as a predictor of overall satisfaction in the first models, and potential for career advancement because we initially hypothesized that recent graduates may take the first job offer they receive, which could limit the individual's potential for further advancement opportunities if they are not fully invested in their position.

For both of these outcomes, we chose to just run the tuned random forest model, given that it was the most effective model in the original analysis. The results from the career advancement model followed a similar pattern as the original analyses, though with lower accuracy. The results from the salary satisfaction analysis are more interesting.


salary
age
gender_M
employer_size_X8
highest_degree_X2
employer_sector_X21
race_X5
employer_size_X3
employer_size_X4
loan_amount_X04
employer_sector_X32
race_X4
race_X3
employer_region_X05
highest_degree_X4
employer_region_X09
employer_region_X02
loan_amount_X05
employer_region_X03
loan_amount_X03
employer_size_X2
employer_sector_X23
loan_amount_X06
employer_size_X6
employer_size_X7
loan_amount_X07
employer_sector_X12
employer_region_X07
employer_region_X08
race_X7
loan_amount_X12
loan_amount_X08
employer_region_X04
employer_size_X5
highest_degree_X3
employer_sector_X31
employer_region_X06
loan_amount_X10
loan_amount_X09
employer_sector_X22
loan_amount_X11
race_X2
race_X6
0
20
40
60
80
100
Importance

Conclusion