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

Connor Harrison, Dong Hoon Lee, Haorui Sun

April 28, 2019

Problem Statement and Background

One of the most prominent issues in contemporary American society is the rising cost of higher education. Between 1988 and 2018, the average cost of attending a 4-year public university has increased from $3,190 to $9,970, an increase of 213 %. This trend is similar, though a smaller magnitude increase, for private universities, at 129%. However, this upward trend has not been accompanied by an upward increase in incomes; CPI adjusted incomes have increased by 13% over this same period.

This disparity between stagnating incomes and rising tuition costs has put higher education out of reach for many middle class families through their own financing. Conversely, the demand for higher education has grown dramatically as the financial and social benefits incurred from possession a bachelor's degree have become more salient in the public eye. To facilitate the pursuit of higher education for children from middle class and low-income families, the U.S. Department of Education has begun offering grants and subsidized loans to students to make up the financing gap families cannot meet on their own.

This issue is not unknown; significant reporting has been done on the growing proportion of student loan debt to total debt held by U.S. families. The total student loan amount has reached $1.5 trillion in 2018 and this makes up 36% of non-housing debt, up from 24% in 2008. The average student from graduating class of 2016 has about $37,000 in student debt. And among 47 million Americans with student loan debt, 11.5% of student loans are 90 days or more delinquent or are in default. These numbers are worrisome by themselves, but the impacts of these loans on the life decisions are even more troubling.

Few studies have reported the downstream effects that this level of debt has on students' post-graduation decision-making. Some of the decisions affected by student debt include first home purchases, marriage and relationship patterns, and career decisions. For instance, students that graduate with significant student loan debt may feel pressure to take the first employment opportunity that they find because students must begin repaying their loans after they graduate. Given this shortened timeline, students' skills may be the best match to the position they acquire, resulting in a loss of employment efficiency. If easily acquired positions are not as high paying as jobs that students could acquire given a longer search timeline, students' long-term earnings could be negatively affected, given that beginning salary is a strong indicator of lifelong earnings potential. Given these insights, we hypothesize that high levels of student loan debt have negative effects on employment outcomes for recent graduates; facing looming debt payments, new job market entrants may spend less time searching for the "right fit" job and instead be pressured into accepting an opportunity that is a less-than-perfect fit.

Research Design and Approach

We used the National Survey of College Graduates (NSCG), a biennial survey that gathers information on college degree field and post-graduation employment outcomes, to examine the the relationships between the student loan debt amount and job mismatch. However, there was no variable in this dataset that directly measure job match or mismatch and, therefore, we used the variable on job satisfaction as an outcome measure for job mismatch for college degree holders under age of 36.

For the analysis, we began by selecting the variables of interest and re-coding them to fit our analysis methods. These variables we used to examine the relationships with the outcome measure are: undergraduate loan amounts, race, age, gender, salary, employer's region, employee size, highest degree attained, and job sector.

To answer the question of whether higher levels of students loan debt affect job satisfaction of college graduates, we used data visualizations to examine and test any trends or relationships between variables of interest. In terms of job satisfaction, we expected see whether those with little to none student debt have higher overall job satisfaction under the assumption that they are able to find jobs with better fits. Conversely, we also expected to see those with higher levels of student loan debt to have lower overall job satisfaction since they were unable to seek for better fit jobs. We also included satisfaction variables for different aspects of the job to compare with between overall satisfaction. With the machine learning analysis, our goal was to find the best predictor for the overall job satisfaction. In addition, we wanted to test whether the level of student loan amounts is one the stronger predictors of the outcome.

Methods

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

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.

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.

Figures 4 through 7 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 the 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 low levels of satisfaction 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. On the other hand, we also observed abnoramlly low levels of satisfaction found in some groups. For example, those with loan amounts of greater than $60,000 had one of the lowest career advancement satisfactions of around 50%. Also, self-employed workers reported unusually low satisfaction level in job security, whereas those working very small companies (less than 10 employees) showed very low satisfaction levels in benefits but high levels in responsibility. These cases suggested that characteristical differecnes of a job can lead to low satisfactions in only one or two aspects of the job and revealed limitations of understanding satisfaction and right fit of a job using a single measurement outcome.

In figures 9, we created bar charts to show the distribution of the responses for job satisfaction and salary satisfaction questionnaires by salary groups. The percent of "satisfied"" and "very satisfied"" responses increased with increasing salary amount. However, the changes were very small for the overall job satisfaction questionnaire and, therefore, difficult to be claimed as meaningful. On the other hand, responses for the salary satisfaction questionnaire showed more variations over the same salary groups and observe that satisfaction increase steeply in the lower salary ranges, but levels out when the salary exceeds $100,000.

Through this visual analysis method, we found few patterns that suggest quadratic relationships between our dependent and independent variables. However, the satisfaction levels were generally very high (around 85%) and had small variations within them, which meant that there was small room for meaningful variations to be observed. And because this analysis is limited to maximum of 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.

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 through the visualizations, we can use machine learning to predict which characteristics of recent graduates influence these outcomes. For this analysis, we have chosen to use three different machine learning models: K-Nearest Neighbor, Regression Tree, and Random Forest. For each model, we will use the base model and a second version that includes tuning parameters. The ouput of all models are included in the appendix.

Which Model predicts our outcome the best?

To compare which of the six machine learning models performed the best on the training data, we created a dotplot after all the models were run on the training data. The dot plot 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 shows that the tuned random forest model has the highest predictive power of the six models. However, the model's specificity is lower than that of the regression tree models. As the ROC metric is the primary measure of interest, we will first test the tuned random forest model on the test data. We will also test the turned regression tree and K-nearest neighbor models to examine any potential deviations in outcome from the random forest model.

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.

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 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 employment in the ______sector, 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 10th 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. Salary and age have a positive relationship to the outcome; while age shows a linear trend, both increase in predictive power as the value of the variable increases. Employment in the ______ sector is the outlier, which has a negative relationship to predicted job satisfaction.

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 87.4% 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 and 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.

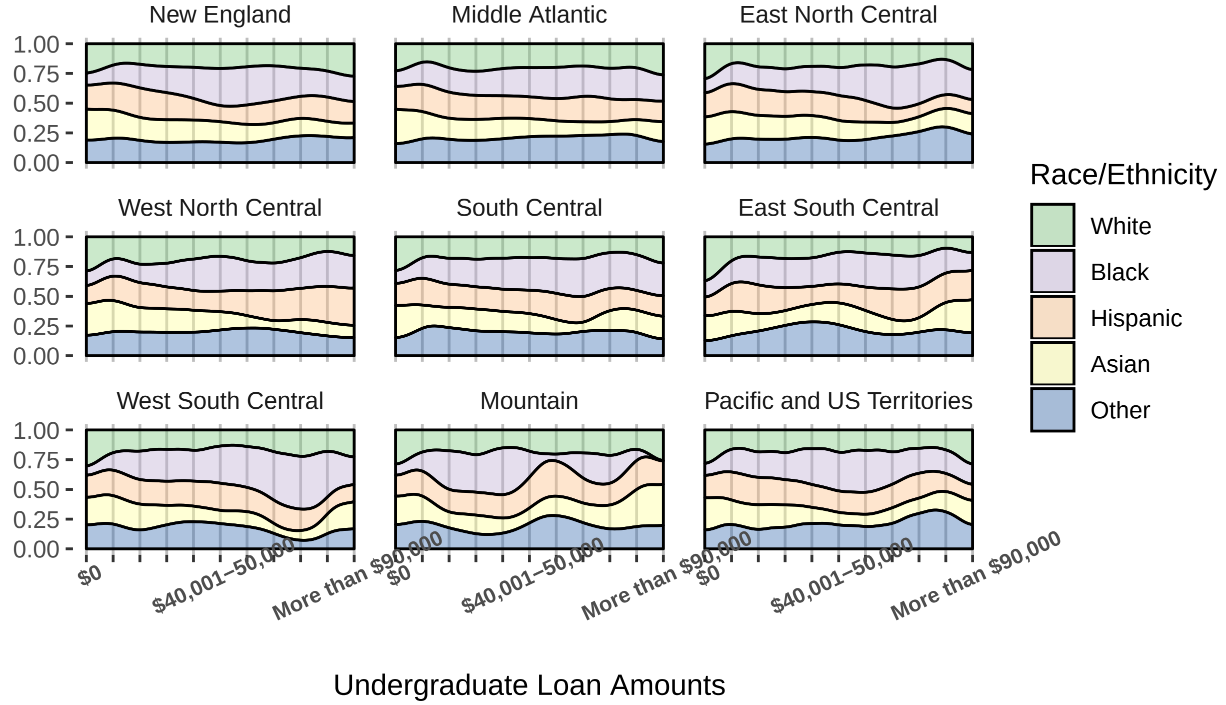
For both of these outcomes, we chose to just run the tuned random forest model, given that it was the most effective model in the original analysis. The results from the career advancement model followed a similar pattern as the original analyses, though with lower accuracy (67.9%). The one important finding of note in this model is that the third strongest predictor of statisfaction with career advancement potential is being male. The results from the salary satisfaction analysis, however, are more interesting. As with the other models, this model still leans heavily towards predicting the positive outcome with a specificity of 95.81%. Interestingly, this model is much better at predicting the true negative than any of the other models, with specificity of 14.1%. However, this does come with a tradeoff in accuracy (76.1%).

Conclusion

This analysis was designed to examine if the data available in the National Survey of College Graduates could be used to predict job mismatch through analyzing individuals' self-reported satisfaction with their current employment. This analysis showed that with the data available, it is quite easy to predict if an individual is "satisfied" with their employment, but this is primarily due to the fact that the vast majority of respondents in the sample reported some level of satisfaction with their employment situation. The number of individuals reported being satisfied with their employment seems to be too high to match reality. We think this could be for two reasons. First, there is a potential response bias, whereby individuals feel pressured to give a positive (or at least not negative) response to an official government survey. Second, there could be a selection issue in respondents; the US economy and job market were strong at the time the survey was implemented, which could indicate that many individuals who graduated around this time were able to find and secure gainful employment. The key insight gleaned from this analysis is how strong of a predictor salary is to job satisfaction. Salary vastly outpaced all other predictive factors, including age, race, region of employment, company size, and employment sector. Second, individuals are more satisfied with their employment as they age, which likely indicates that college graduates are able to successfully sort into better roles as they spend more time in the labor force.

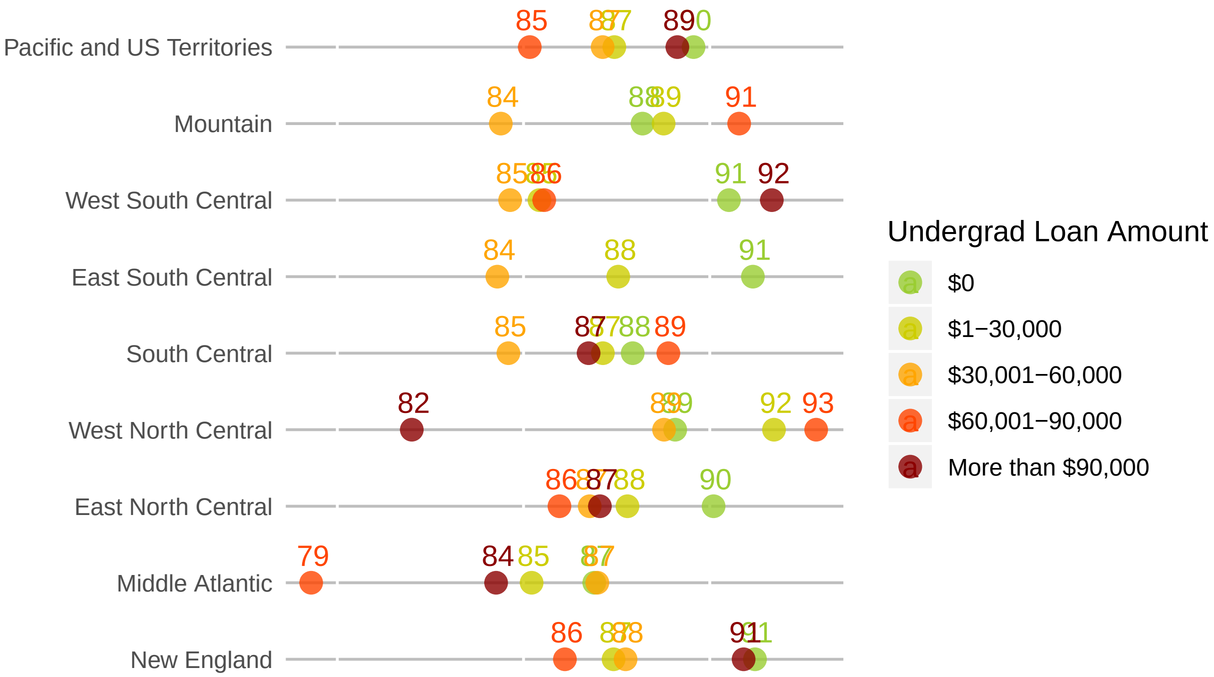
Appendix: Visualizations and Machine Learning Output

Figure 1. Undergrad Loan Borrowing Trend by
Employer Region and Respondent's Race



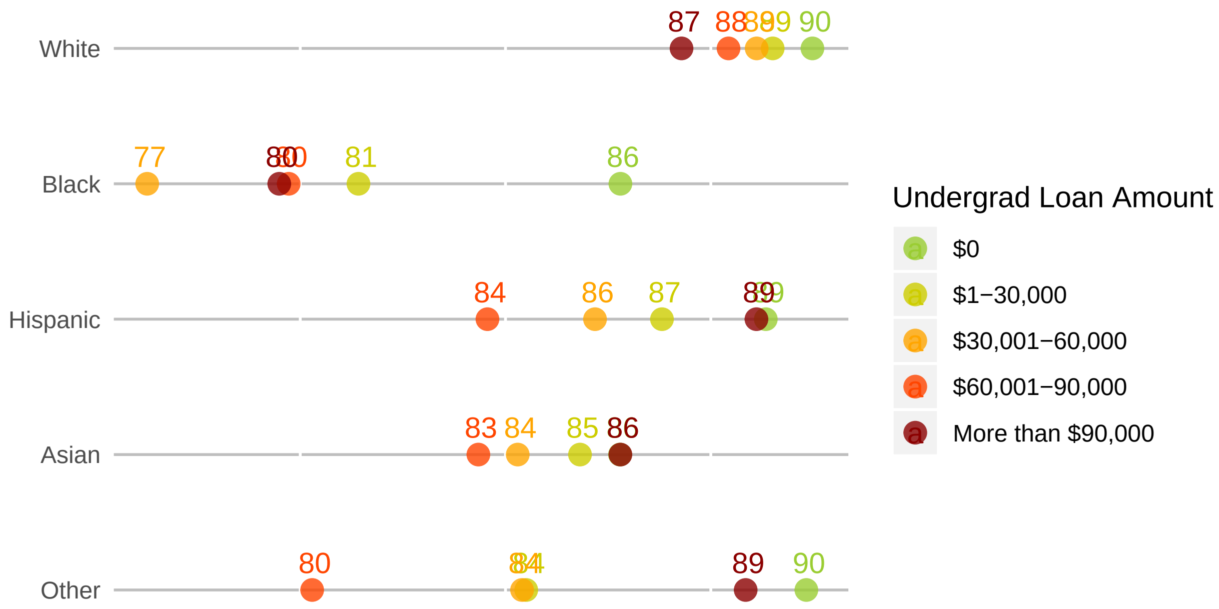
Data: National Survey of College Graduates

Figure 2. Percent Satisfied with Job by Undergraduate Loan Amount and Region of Employer



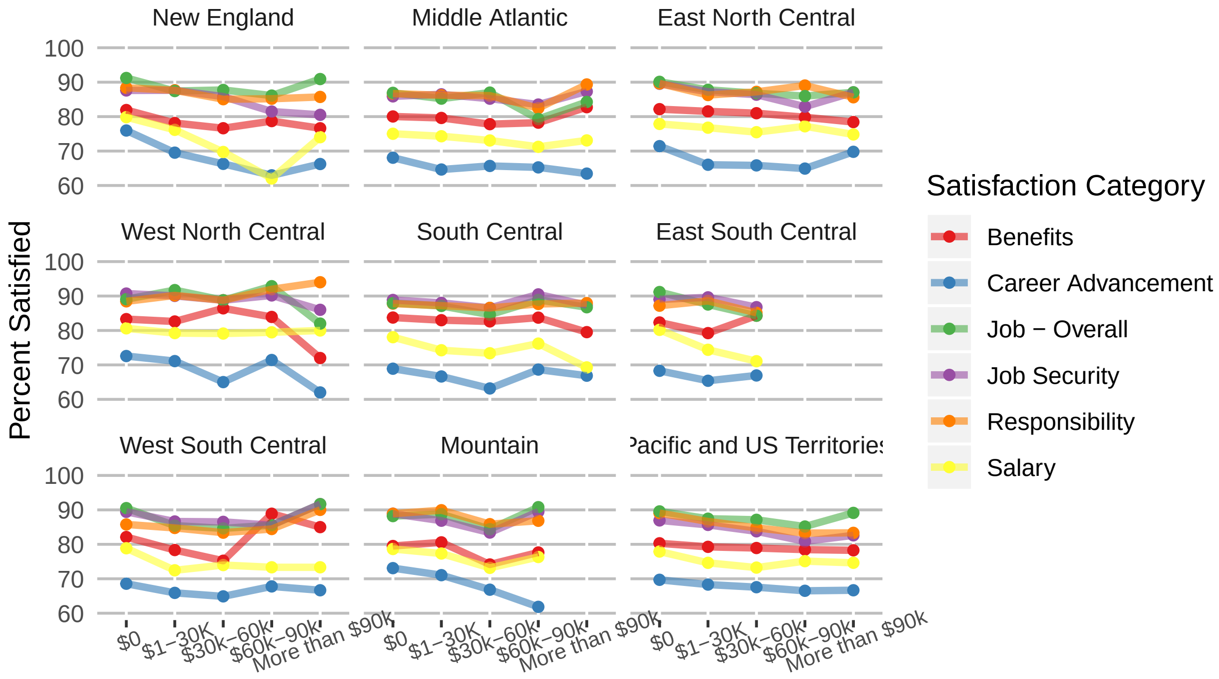
Data: National Survey of College Graduates

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



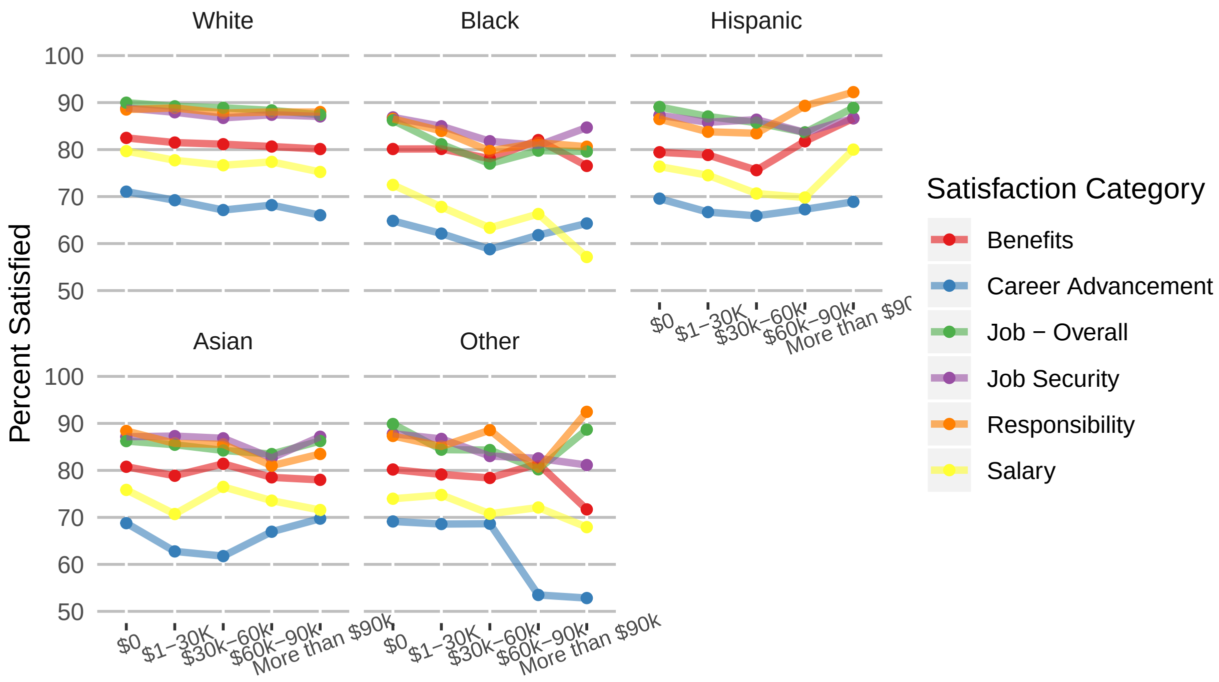
Data: National Survey of College Graduates

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



Data: National Survey of College Graduates

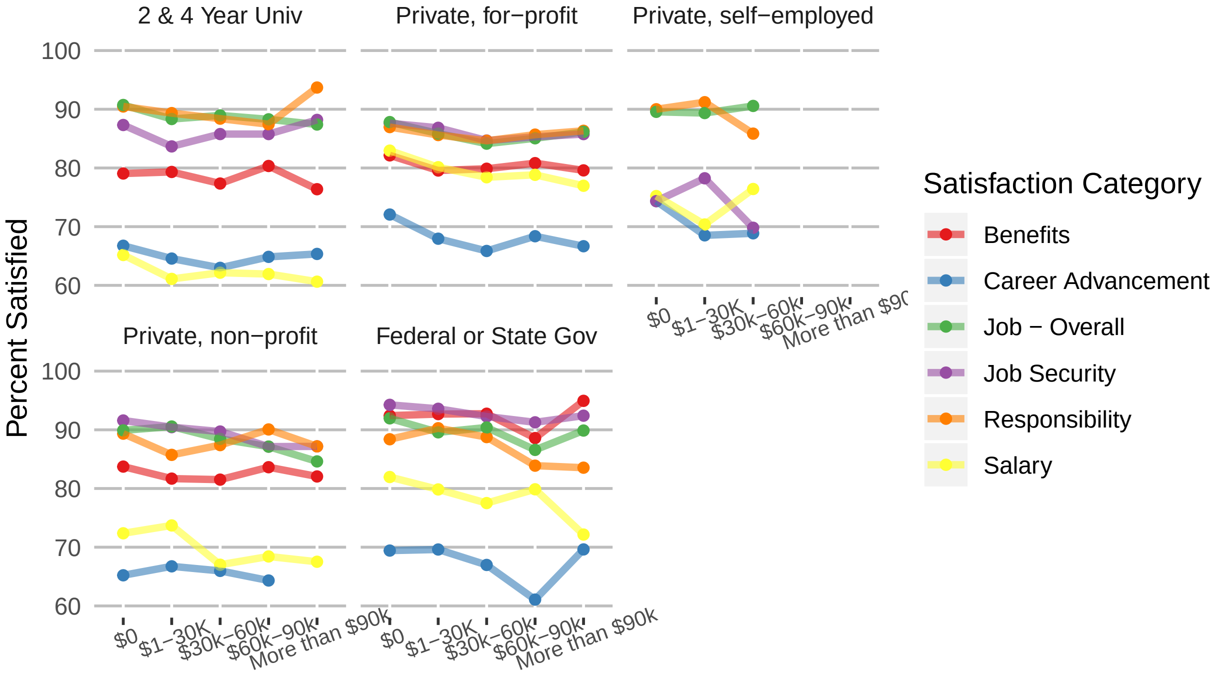
Figure 5. Types of Job Satisfaction by Undergraduate Loan Amounts and by Race



Data: National Survey of College Graduates

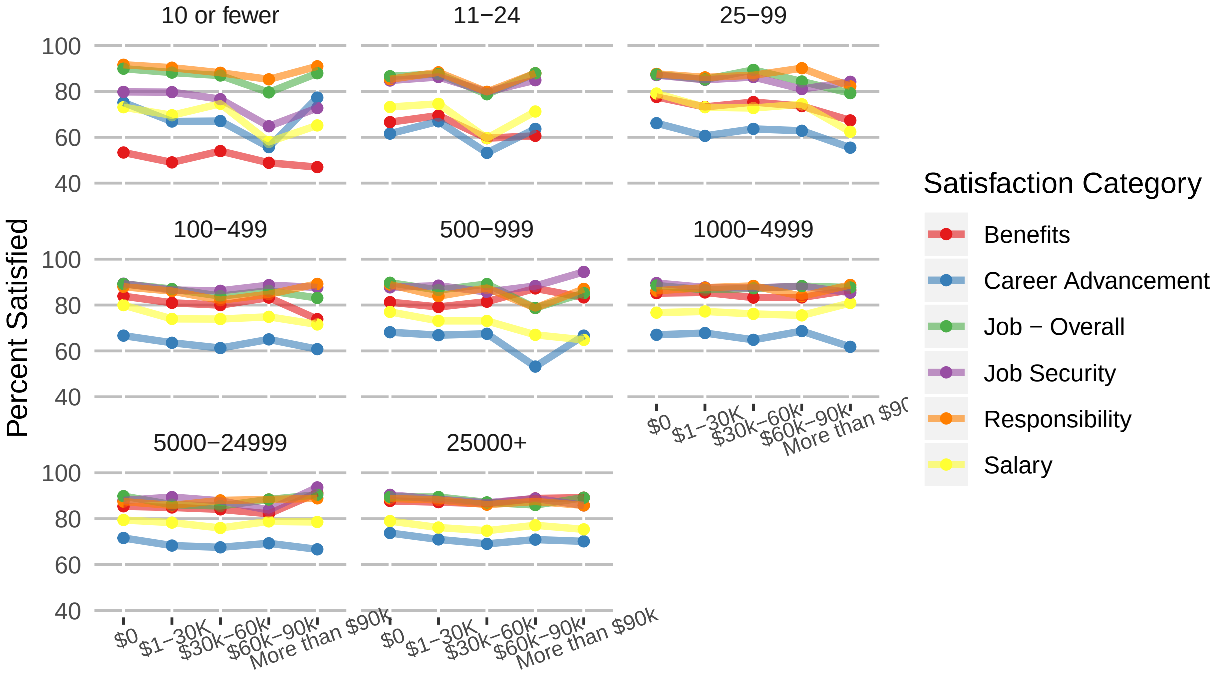
```
## Warning: Removed 4 rows containing missing values (geom_point).
```

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



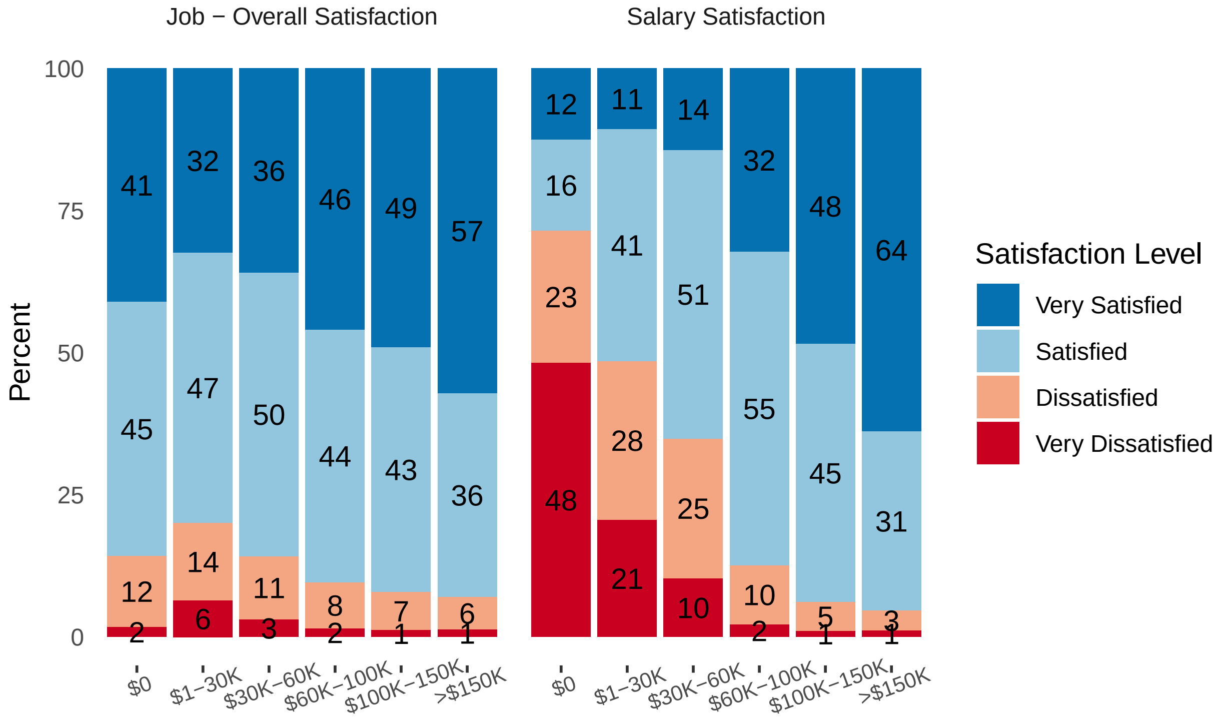
Data: National Survey of College Graduates

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



Data: National Survey of College Graduates

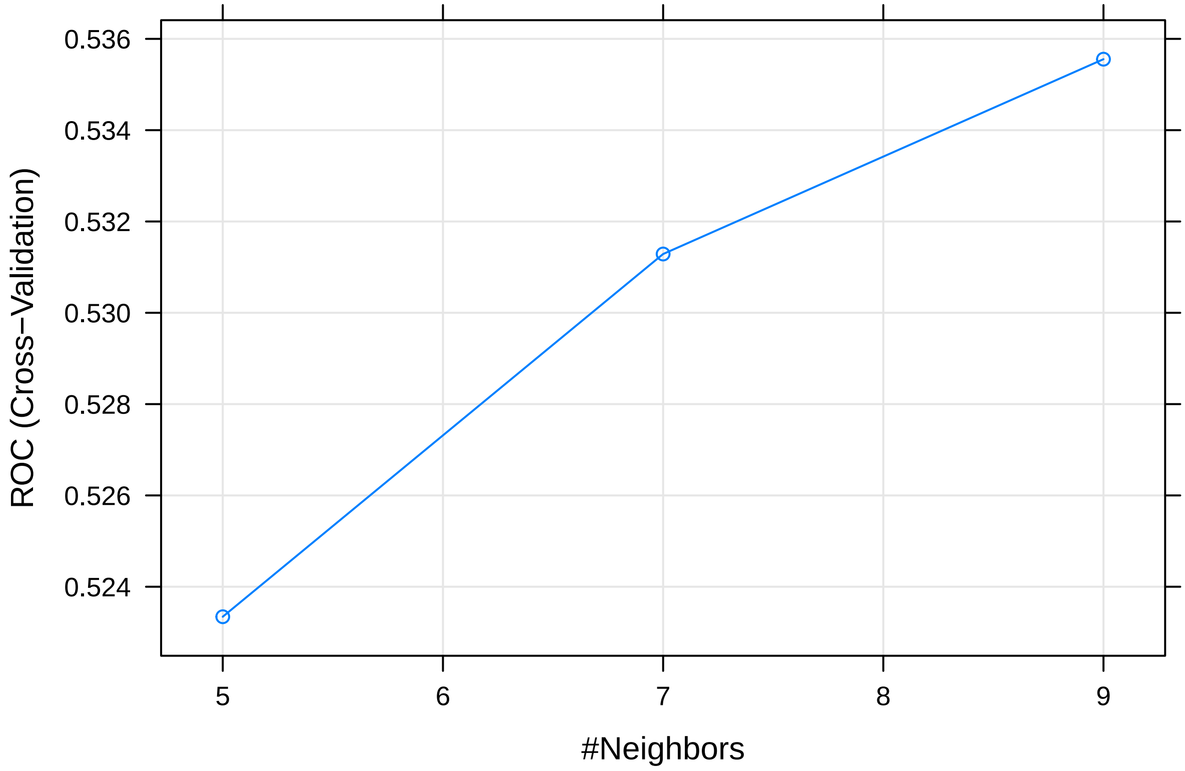
Figure 9. Job and Salary Satisfaction by Salary Groups



Data: National Survey of College Graduates

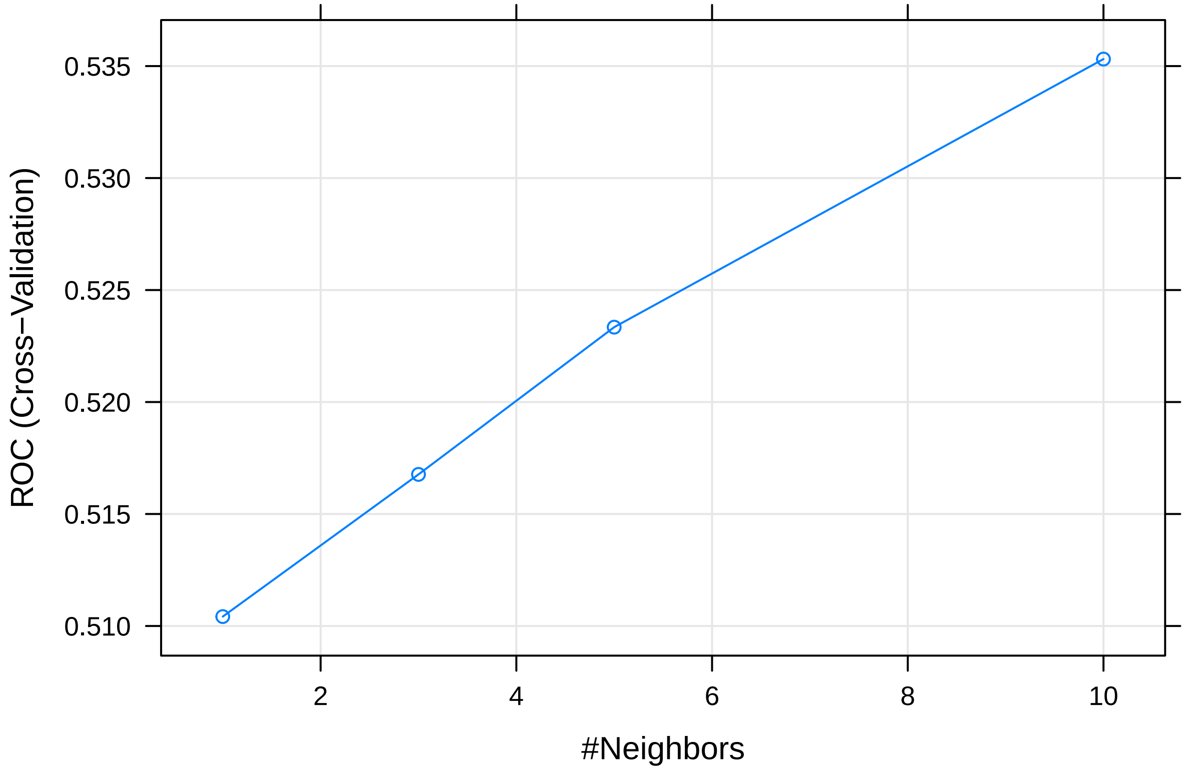
K-Nearest Neighbor Model

```
## k-Nearest Neighbors
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results across tuning parameters:
##
##   k  ROC        Sens       Spec
##   5  0.5233428  0.9831650  0.025220419
##   7  0.5312876  0.9928537  0.012976845
##   9  0.5355544  0.9969079  0.004897085
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```



Tuned K-Nearest Neighbor Model

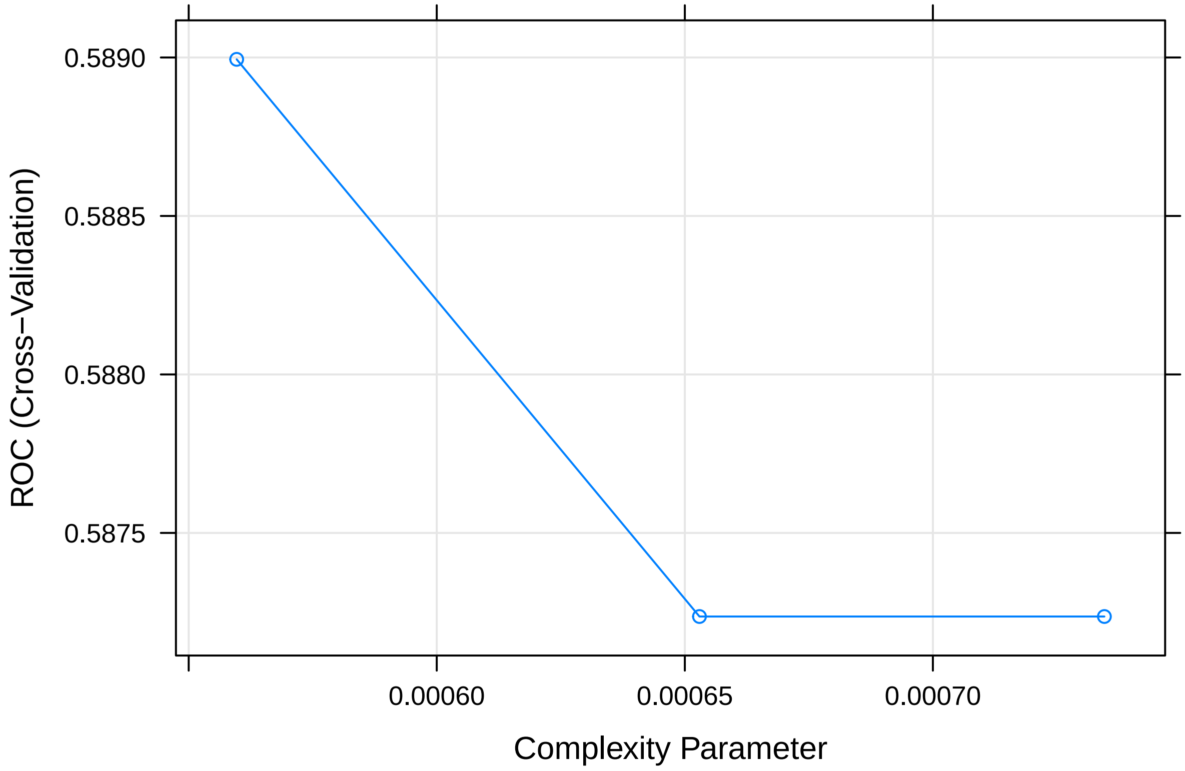
```
## k-Nearest Neighbors
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results across tuning parameters:
##
##   k   ROC        Sens       Spec
##    1  0.5104191  0.8860717  0.134180882
##    3  0.5167680  0.9585137  0.048604179
##    5  0.5233428  0.9831135  0.026077362
##   10  0.5353125  0.9973202  0.004162991
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 10.
```



Regression Tree Model

```
## CART
##
```

```
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results across tuning parameters:
##
##   cp            ROC        Sens       Spec
##   0.0005596754  0.5889941  0.9679102  0.06561854
##   0.0006529546  0.5872362  0.9751252  0.05974309
##   0.0007345739  0.5872362  0.9751252  0.05974309
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.0005596754.
```



Tuned Regression Tree Model

```
## CART
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results:
##
##   ROC        Sens       Spec
##   0.5935329  0.9816015  0.04994854
##
## Tuning parameter 'cp' was held constant at a value of 0.0010281
```

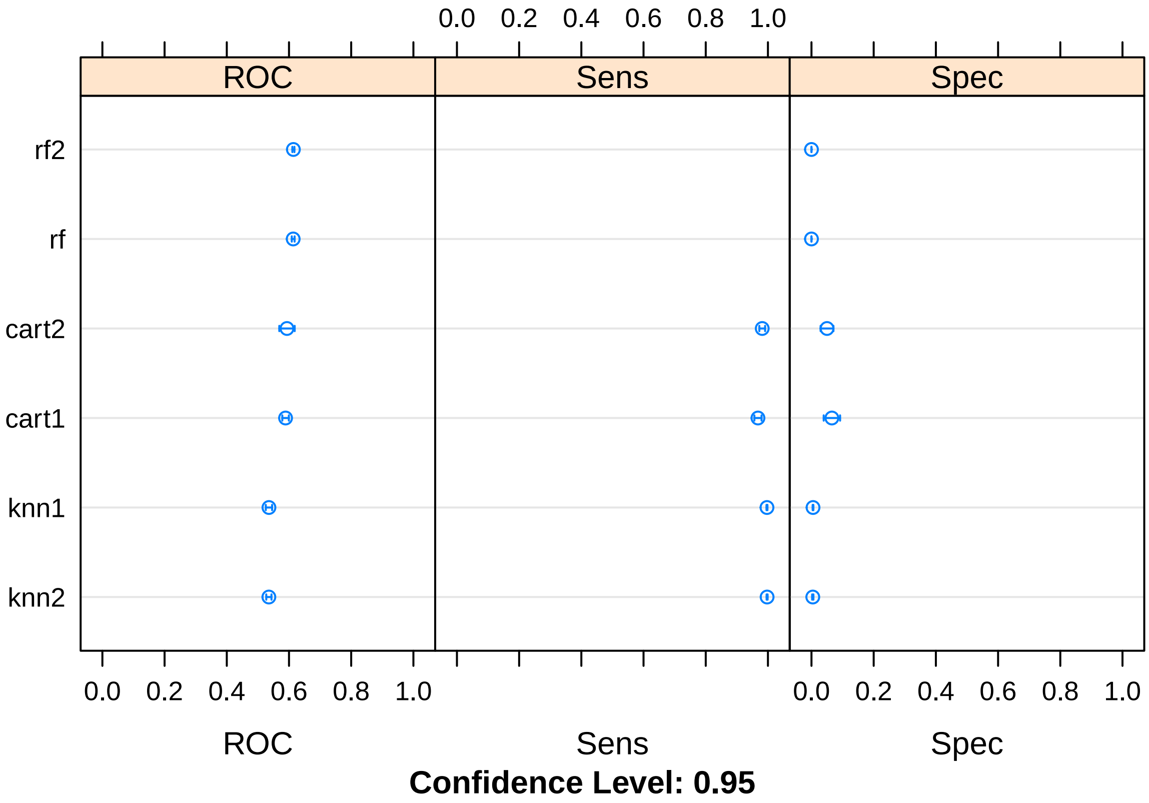
Random Forest Model

```
## Random Forest
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results across tuning parameters:
##
##   mtry  splitrule   ROC        Sens       Spec
##    2    gini        0.6134800  1.0000000  0.00000000
##    2    extratrees  0.5831644  1.0000000  0.00000000
##   22    gini        0.5926630  0.9926991  0.01958722
##   22    extratrees  0.5773009  0.9883357  0.02154650
##   43    gini        0.5879414  0.9877173  0.03158577
##   43    extratrees  0.5808941  0.9848141  0.02766773
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 2, splitrule = gini
##  and min.node.size = 1.
```

Tuned Random Forest Model

```
## Random Forest
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3320, 3319
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    1    0.6136204  1.0000000  0.0000000000
##    2    0.6141517  1.0000000  0.0000000000
##    3    0.6111774  1.0000000  0.0000000000
##    4    0.6059068  1.0000000  0.0006119951
##    5    0.6051752  0.9998110  0.0011016661
##    6    0.6041934  0.9996908  0.0019585342
##    7    0.6036363  0.9992957  0.0026930033
##    8    0.6028095  0.9987975  0.0040398423
##    9    0.6022398  0.9986773  0.0055090053
##   10    0.6019760  0.9981619  0.0057537284
##
## Tuning parameter 'splitrule' was held constant at a value of gini
##
## Tuning parameter 'min.node.size' was held constant at a value of 5
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 2, splitrule = gini
##  and min.node.size = 5.
```

Dot Plot of Machine Learning Model Performance

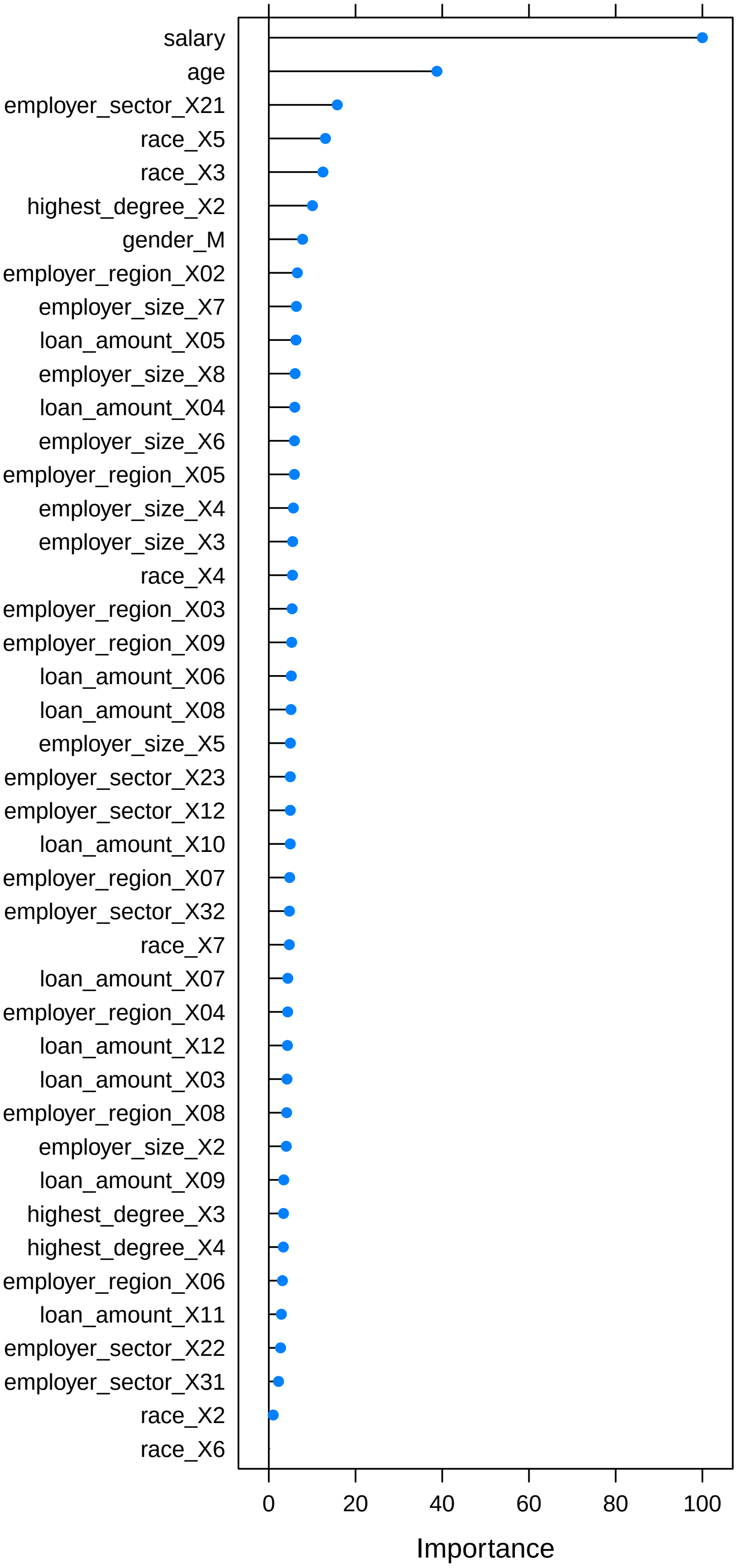


Confusion Matrix: Tuned Random Forest Model

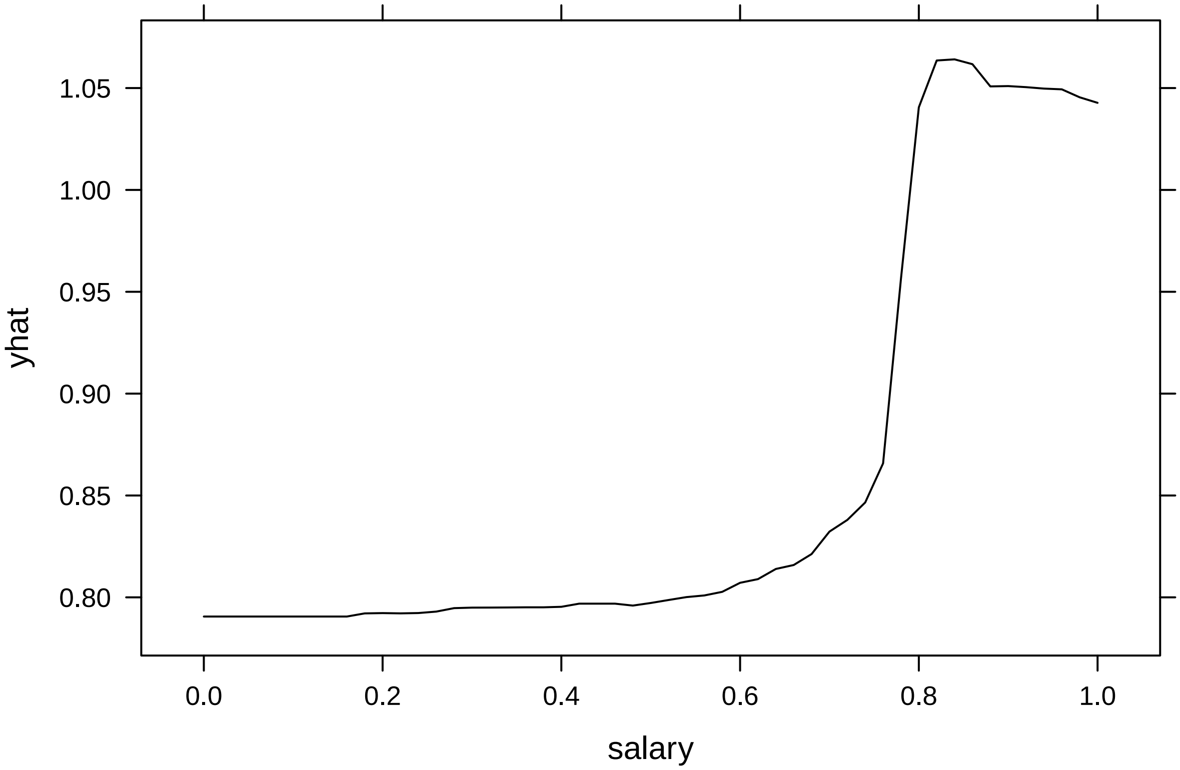
```
## Confusion Matrix and Statistics
##
```

```
##
## 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
##
```

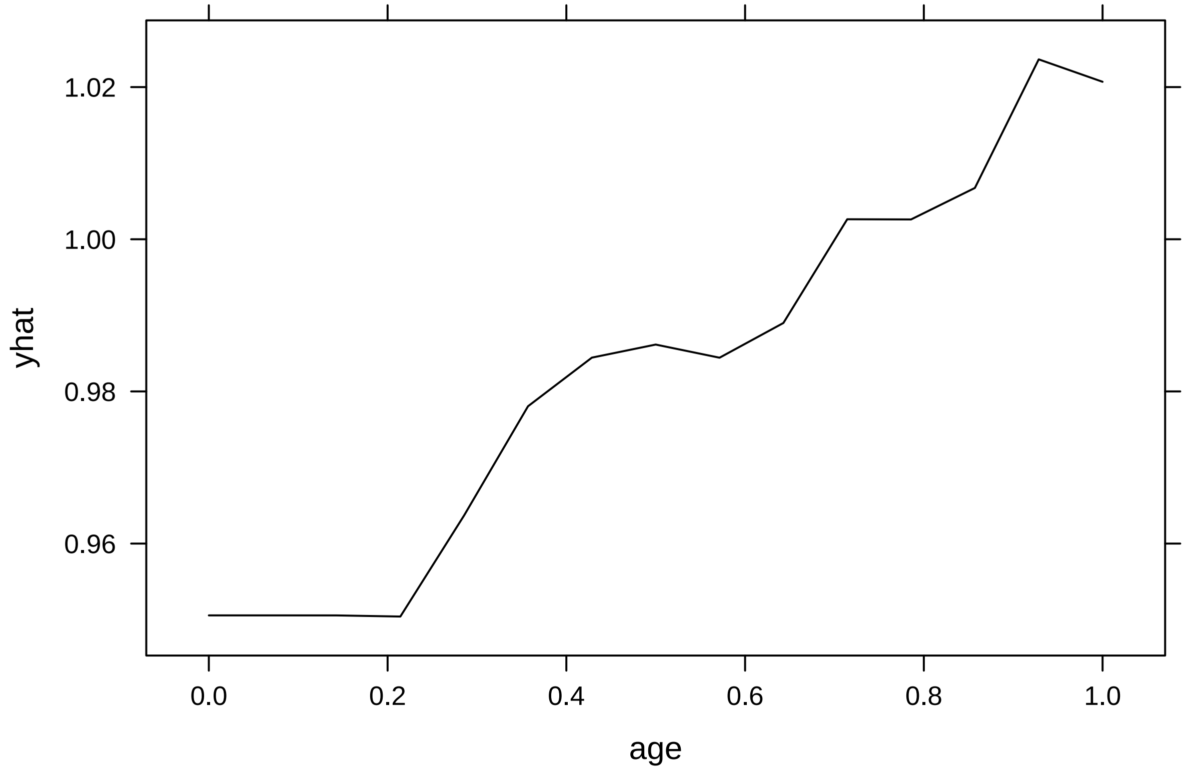
Variable Importance: Tuned Random Forest Model



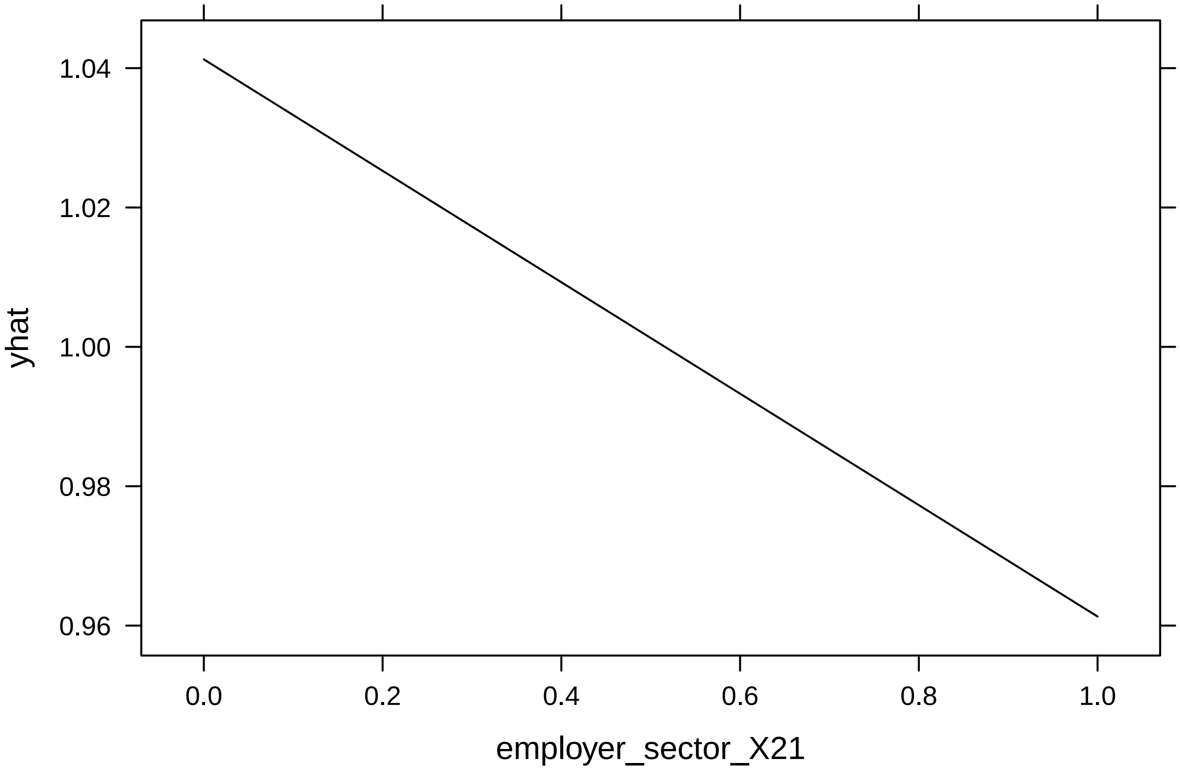
Partial Dependence Plot: Salary



Partial Dependence Plot: Age



Partial Dependence Plot: Employment Sector X_21



Confusion Matrix: Tuned K-Nearest Neighbor Model

```
## Confusion Matrix and Statistics
##
##
## pred            Satisfied Not_Satisfied
##   Satisfied          4838           667
##   Not_Satisfied        13            13
##
##                Accuracy : 0.8771
##                  95% CI : (0.8681, 0.8856)
##     No Information Rate : 0.8771
##     P-Value [Acc > NIR] : 0.5102
##
```

```
##                   Kappa : 0.028
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99732
##             Specificity : 0.01912
##          Pos Pred Value : 0.87884
##          Neg Pred Value : 0.50000
##              Prevalence : 0.87706
##          Detection Rate : 0.87471
##    Detection Prevalence : 0.99530
##       Balanced Accuracy : 0.50822
##
##        'Positive' Class : Satisfied
##
```

Confusion Matrix: Tuned Regression Tree Model

```
## Confusion Matrix and Statistics
##
##
## pred_cart2      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
##
```

Tuned Random Forest Model: Full Sample

```
## Random Forest
##
## 50322 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 10064, 10065, 10064, 10064, 10065
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    1    0.6192364  1.0000000  0.0000000000
##    2    0.6190632  1.0000000  0.0000000000
##    3    0.6146047  1.0000000  0.0000000000
##    4    0.6106014  0.9999945  0.0000000000
##    5    0.6074782  0.9999890  0.0002501126
##    6    0.6054071  0.9999448  0.0004501751
##    7    0.6049849  0.9998787  0.0008003126
##    8    0.6030816  0.9998014  0.0012505377
##    9    0.6027557  0.9996580  0.0018007378
##   10    0.6024076  0.9995587  0.0019008004
##
## Tuning parameter 'splitrule' was held constant at a value of gini
##
## Tuning parameter 'min.node.size' was held constant at a value of 5
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 1, splitrule = gini
##  and min.node.size = 5.
```

Confusion Matrix: Tuned Random Forest Model, Full Sample

```
## Confusion Matrix and Statistics
##
##
## pred_rf2_all    Satisfied Not_Satisfied
##   Satisfied         15107          1665
##   Not_Satisfied         0             0
##
##                Accuracy : 0.9007
##                  95% CI : (0.8961, 0.9052)
##     No Information Rate : 0.9007
##     P-Value [Acc > NIR] : 0.5065
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9007
##          Neg Pred Value :    NaN
##              Prevalence : 0.9007
##          Detection Rate : 0.9007
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Satisfied
##
```

Tuned Random Forest Model: Salary Satisfaction

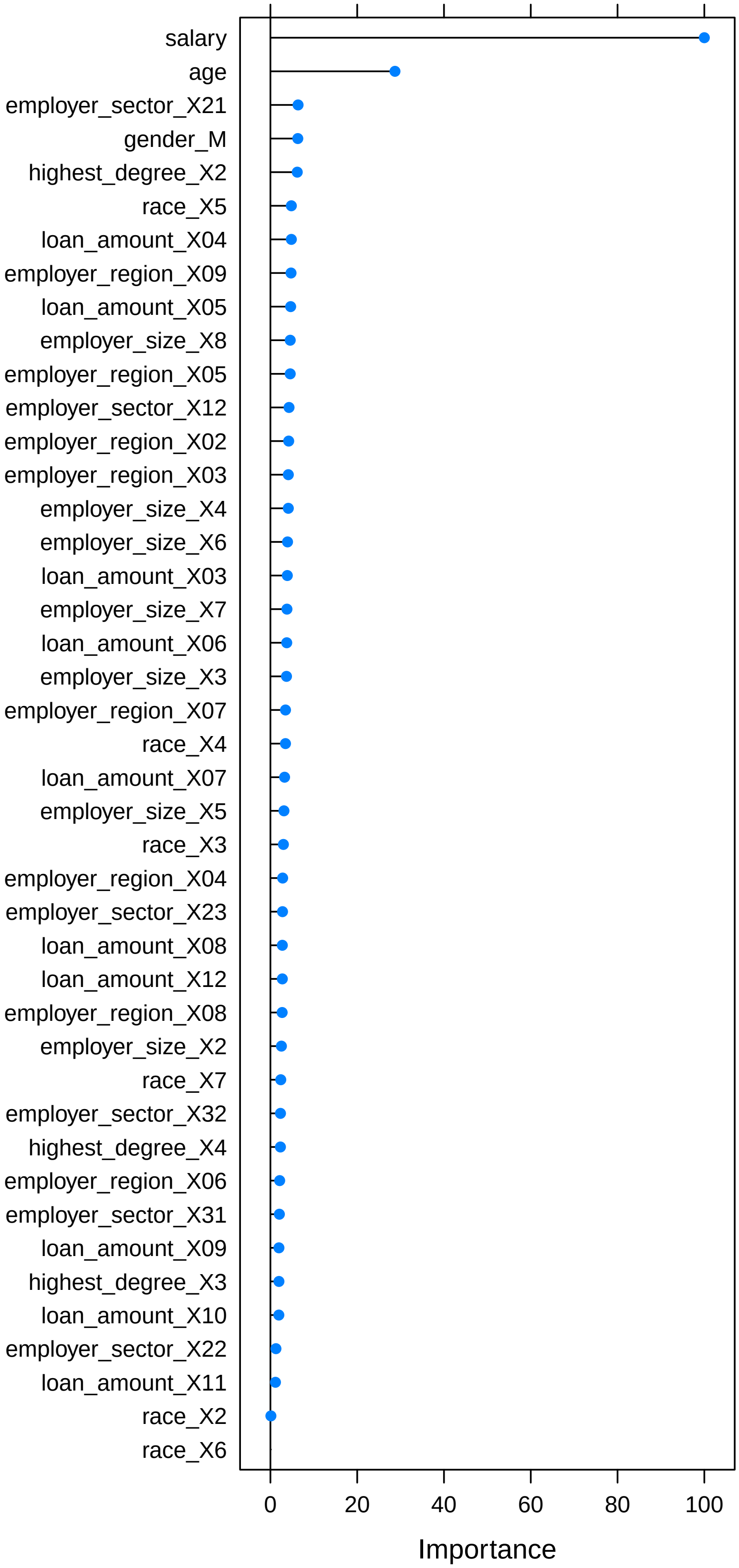
```
## Random Forest
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3319, 3320
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    1    0.7156168  1.0000000  0.0000000000
##    2    0.7249399  0.9999008  0.0006876758
##    3    0.7286722  0.9904533  0.0308827173
##    4    0.7292174  0.9737218  0.0885835222
##    5    0.7293835  0.9593720  0.1334071780
##    6    0.7291196  0.9502620  0.1640398757
##    7    0.7291360  0.9435138  0.1832944866
##    8    0.7288741  0.9379565  0.1994859722
##    9    0.7283853  0.9342450  0.2104263246
##   10    0.7279891  0.9315258  0.2195539036
```

```
##
## Tuning parameter 'splitrule' was held constant at a value of gini
##
## Tuning parameter 'min.node.size' was held constant at a value of 5
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 5, splitrule = gini
##  and min.node.size = 5.
```

Confusion Matrix: Tuned Random Forest Model, Salary Satisfaction

```
## Confusion Matrix and Statistics
##
##
## pred_rf2_salary Satisfied Not_Satisfied
##   Satisfied          4022          1145
##   Not_Satisfied       176           188
##
##                Accuracy : 0.7612
##                  95% CI : (0.7497, 0.7724)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : 0.3597
##
##                   Kappa : 0.1318
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9581
##             Specificity : 0.1410
##          Pos Pred Value : 0.7784
##          Neg Pred Value : 0.5165
##              Prevalence : 0.7590
##          Detection Rate : 0.7272
##    Detection Prevalence : 0.9342
##       Balanced Accuracy : 0.5496
##
##        'Positive' Class : Satisfied
##
```

Variable Importance: Tuned Random Forest Model



Tuned Random Forest Model: Career Advancement

```
## Random Forest
##
## 16595 samples
##    43 predictor
##     2 classes: 'Satisfied', 'Not_Satisfied'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3318, 3319, 3319, 3319, 3320
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    1    0.5920111  1.0000000  0.0000000000
##    2    0.5926387  0.9998891  0.0005161456
##    3    0.5892795  0.9909691  0.0146864649
##    4    0.5873527  0.9723749  0.0482353753
##    5    0.5846397  0.9555116  0.0746525190
##    6    0.5836862  0.9421318  0.0946887721
##    7    0.5827628  0.9351201  0.1077328927
##    8    0.5831397  0.9279089  0.1184782507
##    9    0.5820286  0.9210968  0.1250944994
##   10    0.5820812  0.9187227  0.1279564769
##
## Tuning parameter 'splitrule' was held constant at a value of gini
##
## Tuning parameter 'min.node.size' was held constant at a value of 5
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 2, splitrule = gini
##  and min.node.size = 5.
```

Confusion Matrix: Tuned Random Forest Model, Career Advancement

```
## Confusion Matrix and Statistics
##
##
## pred_rf2_adv    Satisfied Not_Satisfied
##   Satisfied          3755          1776
##   Not_Satisfied         0             0
##
##                Accuracy : 0.6789
##                  95% CI : (0.6664, 0.6912)
##     No Information Rate : 0.6789
##     P-Value [Acc > NIR] : 0.5064
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.6789
##          Neg Pred Value :    NaN
##              Prevalence : 0.6789
##          Detection Rate : 0.6789
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Satisfied
##
```

Variable Importance: Tuned Random Forest Model, Career Advancement

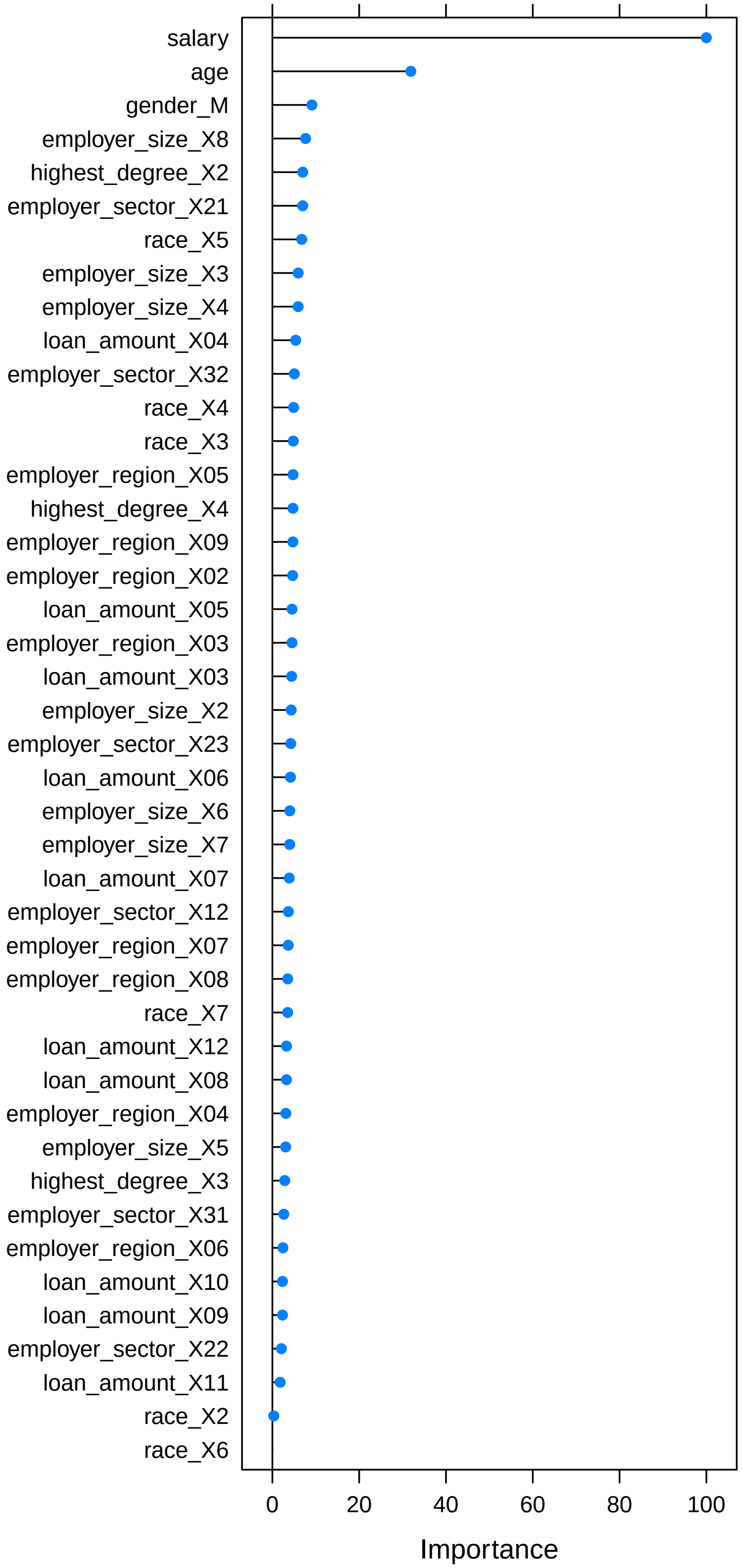
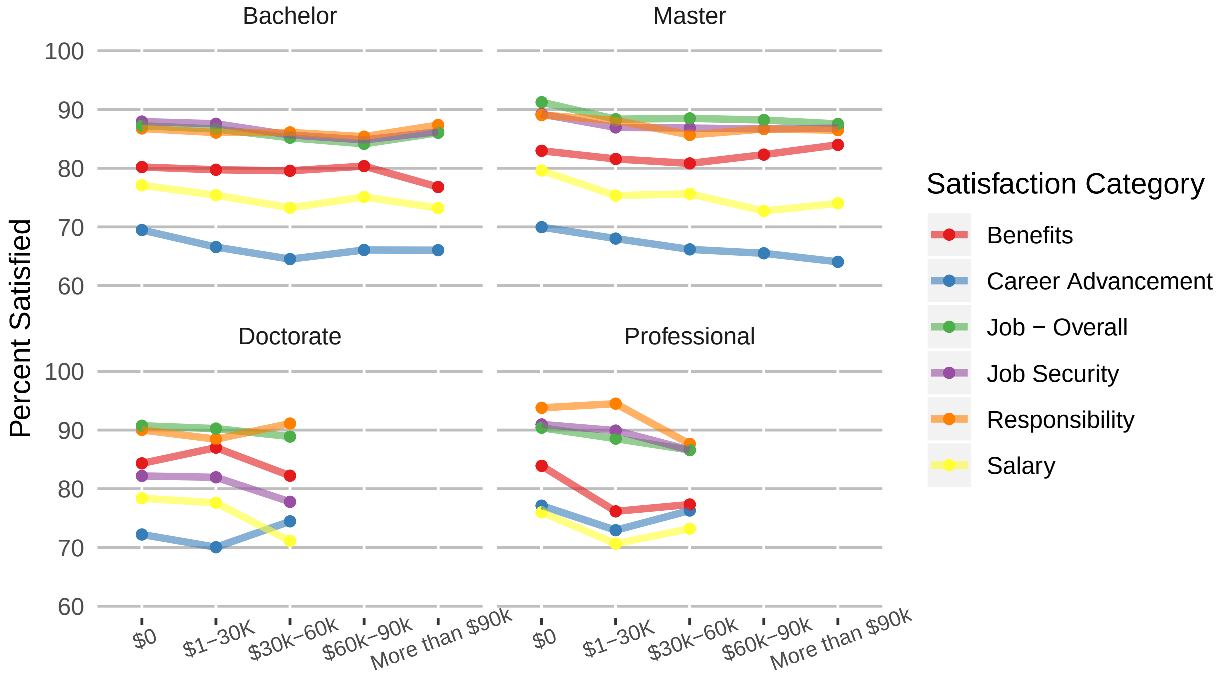
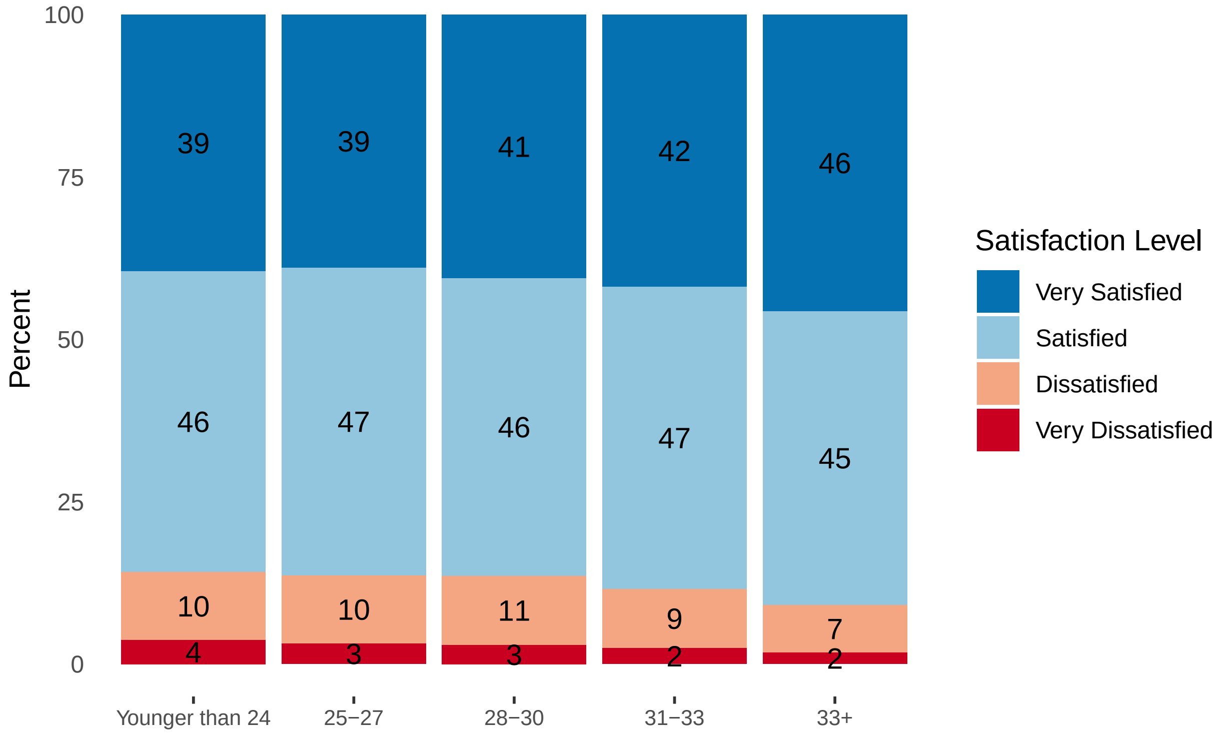


Figure A1. Types of Job Satisfaction by Undergraduate Loan Amounts and Highest Degree



Data: National Survey of College Graduates

Figure A2. Job Satisfaction by Age Groups



Data: National Survey of College Graduates