ORGB 671: Assignment #3
US Patent Office: A Predictive Model

ORGB 671

January 31, 2023

Group 4:
Emery Dittmer 260658030
Hugo Garcia 260791363
Liliana Tretyakova 261086215
William Stephenson 261082056

1.1. US Patent Office

The United States Patent Office (USPTO) USPTO advises the president of the United States, the secretary of commerce, and U.S. government agencies on intellectual property (IP) policy, protection, and enforcement; and promotes the stronger and more effective IP protection around the world according to their website.

The USPTO is a large employer of patent examiners. In this basic study we examined the human metrics of these patent employees. This includes turnover rate, mobility within the company and more. To accomplish this we used a dataset provided by the USPTO (simplified by our instructor). Additional details can be found here and you can access the datasets here.

2.1. Modeling: Data Manipulation & Modeling

In order to test the art unit mobility of the examiners, we first extracted a clean dataset as provided by the course. With this data, we counted the art units throughout an examiner's career, before normalizing that to an annual number. The annual art unit mobility then becomes the number of art units through career divided by the number of tenure days times 365. This, therefore, generates a unique and normalized approach to art unit mobility.

To create a Boolean field, we measured the average mobility of the dataset and considered that any art unit mobility (annually adjusted) above the average was a 1, whereas it would be a 0 otherwise. We preformed other data manipulations, such as removing 'last day worked', table joining, and removals of NAs as needed. First, we evaluated a logistic regression model using the factors in the dataset. This demonstrated that 'tenure days' and 'tc' were the essential features based on their significant codes.

2.2. Modeling: Model Testing

We investigated a simple regression tree using all available data and a common control parameter of 0.01. This gave us insights into which factors were the most important and which might be throwing off the results. Through this process, we understood that the 'end date' would not be a helpful variable for modelling this predictor. The following tree model with the appropriate factors predicted that tenure days were the most important factor of higher-than-average art unit mobility.

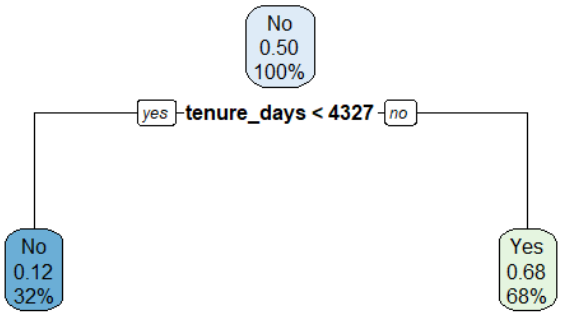

Figure 1: Decision Tree for All Data. "Yes" indicates above average annual Examiner Art Mobility (cp=0.01)

After randomly splitting the data into training and testing data sets, through a 70/30$ split respectively, we evaluated the model performance through accuracy and ROC curves.

Evaluation of Training and Test

```
summary(train)
```

```
##     art_unit       gender            start_year   latest_date
##  Min.   :1600   Length:2762        Min.   :2000   Length:2762
##  1st Qu.:1714   Class :character   1st Qu.:2000   Class :character
##  Median :2111   Mode  :character   Median :2002   Mode  :character
##  Mean   :1976                      Mean   :2003
##  3rd Qu.:2169                      3rd Qu.:2005
##  Max.   :2496                      Max.   :2015
##   tenure_days         tc          work_group   art_unit_distinct_changes
##  Min.   : 267   Min.   :1600   Min.   :1600   Min.   : 1.000
##  1st Qu.:3884   1st Qu.:1700   1st Qu.:1710   1st Qu.: 1.000
##  Median :5194   Median :2100   Median :2110   Median : 2.000
##  Mean   :4750   Mean   :1932   Mean   :1972   Mean   : 2.567
##  3rd Qu.:6218   3rd Qu.:2100   3rd Qu.:2160   3rd Qu.: 3.000
##  Max.   :6518   Max.   :2400   Max.   :2490   Max.   :17.000
##  high_mobility glm_prediction rlm_prediction tree_prediction
##  No :1395      No :1521       No :1521       No :1167
##  Yes:1367      Yes:1241       Yes:1241       Yes:1595
```

Figure 2: Summary of Train Data Set

```
summary(test)
```

```
##     art_unit       gender            start_year   latest_date
##  Min.   :1609   Length:1205        Min.   :2000   Length:1205
##  1st Qu.:1723   Class :character   1st Qu.:2000   Class :character
##  Median :2115   Mode  :character   Median :2002   Mode  :character
##  Mean   :1988                      Mean   :2003
##  3rd Qu.:2174                      3rd Qu.:2005
##  Max.   :2495                      Max.   :2016
##   tenure_days         tc          work_group   art_unit_distinct_changes
##  Min.   :  27   Min.   :1600   Min.   :1600   Min.   : 1.000
##  1st Qu.:3782   1st Qu.:1700   1st Qu.:1720   1st Qu.: 1.000
##  Median :5116   Median :2100   Median :2110   Median : 2.000
##  Mean   :4718   Mean   :1943   Mean   :1984   Mean   : 2.621
##  3rd Qu.:6207   3rd Qu.:2100   3rd Qu.:2170   3rd Qu.: 3.000
##  Max.   :6350   Max.   :2400   Max.   :2490   Max.   :21.000
##  high_mobility glm_prediction rlm_prediction tree_prediction
##  No :607       No :671        No :671        No :508
##  Yes:598       Yes:534        Yes:534        Yes:697
```

Figure 3: Summary of Test Dataset

We then evaluated a logistic regression model, before evaluating a tree model based on the training dataset.

```
Call:
glm(formula = high_mobility ~ art_unit + gender + start_year +
    tenure_days + tc + work_group, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9196  -0.8549  -0.2097   0.8168   2.7688

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.241e+01  6.665e+01   0.786 0.431679
art_unit    -1.651e-02  1.873e-02  -0.882 0.378044
gendermale   5.990e-02  1.004e-01   0.597 0.550786
start_year  -2.872e-02  3.313e-02  -0.867 0.386118
tenure_days  8.378e-04  7.631e-05  10.980  < 2e-16 ***
tc           7.641e-03  1.983e-03   3.853 0.000117 ***
work_group   9.517e-03  1.855e-02   0.513 0.607945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3828.7  on 2761  degrees of freedom
Residual deviance: 2942.9  on 2755  degrees of freedom
AIC: 2956.9

Number of Fisher Scoring iterations: 5
```

Figure 4: Logistic Regression Output and Coefficients

To determine the optimal control parameter (cp) we ran a simple optimal tree computation. This gave an optimal cp value of 0.0015. The optimal cp produced the final tree model shown in figure 7.

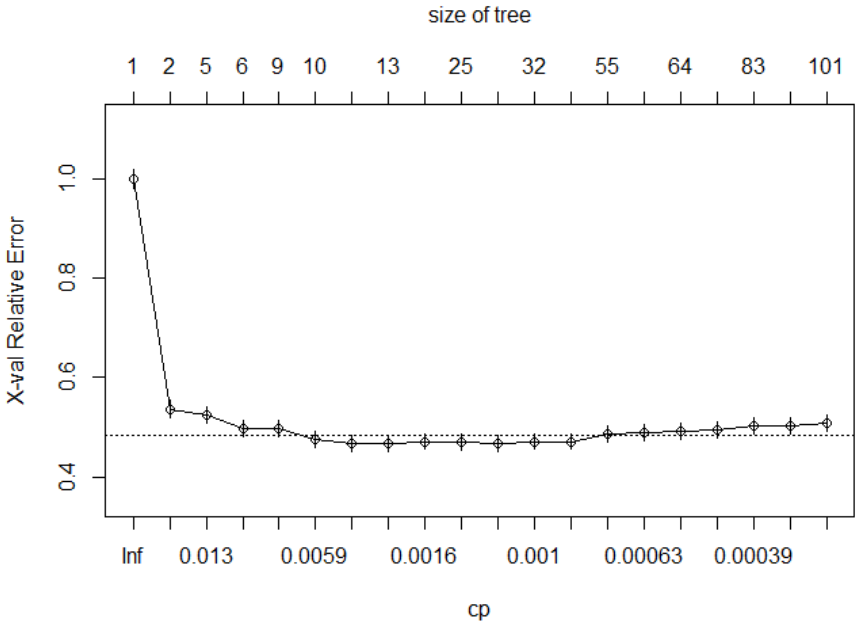

Figure 5: Optimal cp for Decision Tree Using Training Data set.

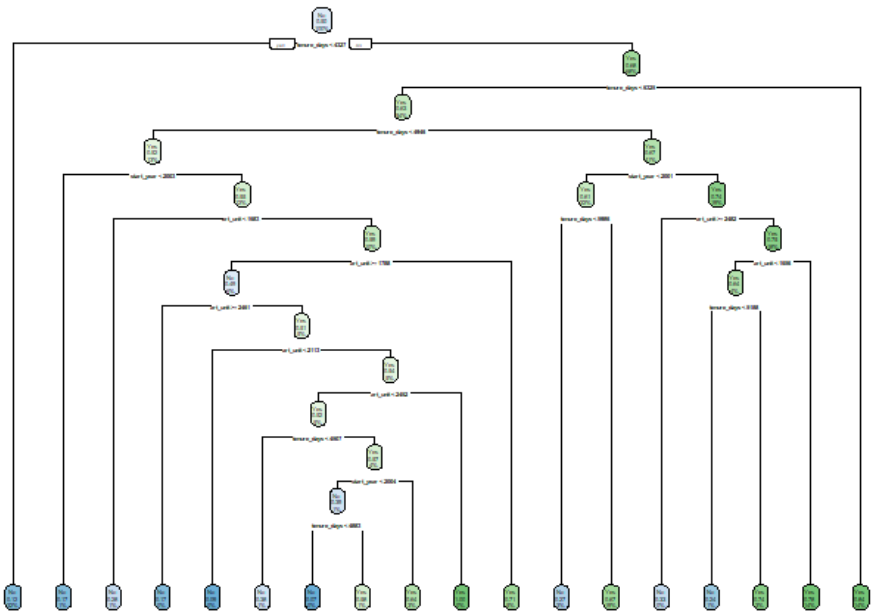

Figure 6: Decision Tree Using Training Data Set (optimal cp=0.0015) "Yes" indicates above average annual Examiner Art Mobility

2.3. Modeling: Evaluation of the Predictive Model

We tested and validated subsets of the data using accuracy, precision and recall.

Comparison of Model Performance.

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.7858921	0.8483034	0.7001647	0.7671480
Logistic Regression	0.7053942	0.6863905	0.7644152	0.7233048

Table 1: Performance Comparisons between models

We then plotted an ROC curve the logistic regression and tree models. The ROC curve, shown in Figures 8 and 9, is a common way to assess the performance of a classification model. In the graphs shown above x-axis represents a false positive rate, and the y-axis represents a true positive rate (both rates are shown from 0 to 1). The perfect model would have a false positive rate of 0 and a true positive rate of 1. Thus, the better the model is, the closer the curve is to the top left corner of the graph. If the model's ROC curve equals the baseline (dotted line), it would indicate that a model is as good as random guessing. Since classes in the given dataset are balanced, the ROC curve can be used to evaluate predictive models.

Figure 8 shows that the logistic regression model is away from the baseline. Its false positive rate is low (approximately 0.2), and true positive rate is relatively high (approximately 0.65). Consequentially, the model's performance is good. However, Figure 9 indicates that the decision tree model performs significantly better. Its ROC curve is far from the baseline, and its true positive rate equals 0.9.

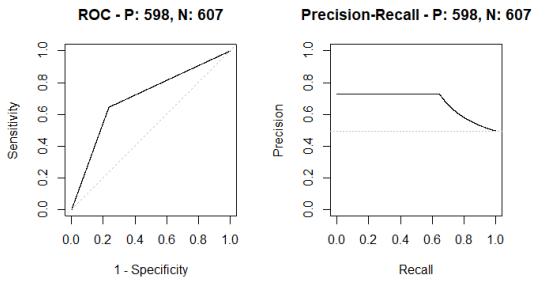

*Figure 7: ROC Curve for **Logistic Regression** with Test Data*

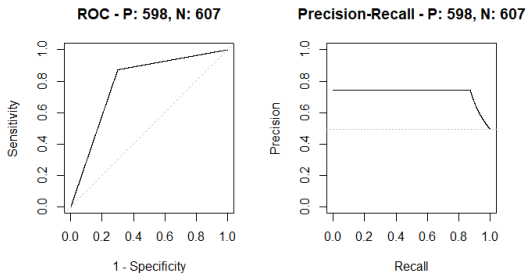

*Figure 8: ROC Curve for **Decision Tree** Model with Test Data*

Another metric that allows us to evaluate the model's accuracy is the Area Under the ROC Curve (AUC). An area of 1 indicates a perfect classifier, and an area of 0.5 represents a model that is no better than random guessing. The AUC of the logistic regression model is 0.70, which is a relatively good result. The decision tree model has an AUC of 0.79. Thus, the decision tree significantly outperforms logistic regression and can facilitate decision-making.

Model	AUC
Logistic	0. 7049501
Tree	0. 7865372

Table 2: AUC model Evaluation

2.4.

Figures 5 and 6 also show precision-recall curves for the two models. Logistic regression and decision tree models appear to have good predictive power as they maintain a high precision and high recall across their graphs. Although, the decision tree precision-recall curve indicates the model performs better than logistic regression.

3.1. Interpretation

The **Decision Tree** as seen in Figure 4 above: this model's output illustrates how examiners are moving between art units over time. This may be due to them reaching the highest point in their career, leave, or moving to another work group or technology center. In any case, it is a significant observation. As a manager, I would ascertain that gender has no or minimal impact on the movement in the company.

The **Logistic Regression** model's output, as described in Figure 1 above and highlighted in Figure 7 below, tells us that the only two meaningful observations are: 'TC' (Technology Center) and 'Tenure Days' (seniority). Therefore, the mobility of examiners has been historically consistent and not differentiable by gender, indicating little to no significant gender bias amongst USPTO examiner mobility across Art Units.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.241e+01  6.665e+01   0.786 0.431679
art_unit    -1.651e-02  1.873e-02  -0.882 0.378044
gendermale   5.990e-02  1.004e-01   0.597 0.550786
start_year  -2.872e-02  3.313e-02  -0.867 0.386118
tenure_days  8.378e-04  7.631e-05  10.980  < 2e-16 ***
tc           7.641e-03  1.983e-03   3.853 0.000117 ***
work_group   9.517e-03  1.855e-02   0.513 0.607945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: Logistic Regression Output and Coefficients

On the significance of 'Tenure Days', although art units and tenure could be colinear, given that we normalized to annual art unit mobility by year, we are able to eliminate this bias.

While the Tree Model is better for accuracy and precision, we care more about the F1 Score as we seek to balance how often we predict correctly so that our predicted coefficients are more useful and reliable for predictions.

3.2. Application

From the models showing a significant relationship for examiner mobility across art units based on tenure days and technology center, some managerial recommendations could be:

- Retention strategies: Focus on retaining examiners with longer tenure and those working in specific technology centers. An examiner with a long tenure is most likely to move across art units. Monitor and track employee movement across art units regularly to identify trends and take proactive measures to ascertain the retention of the most valuable examiners.

- Performance evaluation: Regularly evaluate the performance of employees in relation to their tenure days and technology center and identify areas for improvement. Encourage examiner development and growth opportunities within art units to retain them longer.
- Job rotation: Evaluate the factors contributing to employee movement, such as insufficient career growth opportunities for longer-tenured examiners or poor management practices in a technology center and address them promptly. Consider rotating employees with longer tenure and those working in “stable” technology centers to different departments and technology centers to increase their job satisfaction and keep them engaged, and reduce the risk of unwanted mobility (i.e. leaving altogether).