McGill University

ORGB 671: Talent Analytics

Group Project Report

Presented to:
Prof. Roman Galperin

Arnav Gupta
260658711
Kaz Hayashi
261177330
Chien Chen Liew
261178093
Tony Xu
261137773

Contents

1 Problem Statement and Challenge

The primary challenge of this project lies in understanding and addressing the factors that contribute to the variability in employee attrition rate at the USPTO, with an aim to potentially improve certain procedures within the organization and make the department more attractive for the employees. This involves analyzing organizational and social factors that influence examiner attrition, as well as the impact of gender, race, and ethnicity on the examination process. Identifying patterns of examiner's individual demographics characteristics, inter-department mobility, workload, and attrition among examiners based on demographic characteristics is crucial for developing strategies to examine which affects the employee attrition rate the most. The analysis should offer insights into how the USPTO can implement changes to promote equity among examiners and improve the overall employee satisfaction rate, thereby supporting inventors and the broader innovation ecosystem more effectively. Finally, the team proposed a potential solution (PatentLens) that could address the problems we identified in the analysis.

2 Part 1

2.1 Data pre processing

1. **Feature enrichment:** through the help of packages `gender` and `wru`, we reference the demographics details of the examiner through their first name.
2. **Data cleaning:** Removed records with 'NA' values in the gender field to ensure the dataset only includes entries with known gender.
3. **Feature Engineering:**
   - Added tenure days based upon the minimum date and maximum date for each specific examiner.
   - Added variables related to the quantity of applications in various stages (new, pending, issued, etc.).
   - Added `AU_move_indicator`, a binary variable indicating whether an examiner has moved cross-department in their tenure.
   - Added `Separation_indicator`, a binary indicator suggesting whether an employee has left the firm given the observation date of the dataset.

   At this point, we attempted to aggregate the data by quarter and `examiner_id`, we quickly found out that the examiners could move their art units multiple times within each quarter. This made the data structure unsuitable for panel data analysis. Moreover, as we are trying to see the effect of individual or social characteristics such as gender and race, panel data analysis - fixed effects model was inappropriate. Within fixed effects regression, time-invariant variables are absorbed by the α or entity effects. Therefore, we can only estimate the effect of variables that change over time and among entities (i.e., examiner). Thus, we decided to aggregate the data by `examiner_id` and added variables such as "ratio of woman, minority, their own race" in their unit, enhancing the demographic analysis capabilities and fostering a deeper understanding of diversity within the workforce.

   Finally, the ratio of women, minorities, and individuals of the same race within each unit has also been calculated. This addition allows for an examination of the diversity and representation within units, relating these factors to employee attrition rate.

2.2 EDA

We conducted EDA using R's DataExplore package, which works similar to Python pandas-profiling. The provided dataset consists of 4849 observations of 22 variables related to patent examiners, encompassing identifiers, application process stages, departmental movements, and demographic information. See Figure A.1

The created dataset does not contain any null or missing values(See Figure A.2) However, it is critical in data preprocessing step, we dropped rows that were not identified as either gender based on their names. This could potentially create selection bias. For instance, the data may not contain foreign born examiners because the gender package predicts the gender based solely on the U.S. birth records. If "not being on the record" can affect both turn-over and predictors of our interest, then it might introduce omitted variable bias in our model.

2.2.1 Univariate Distributions

The univariate distribution histograms reveal a diverse range of data characteristics across the variables (See Figure A.3). Abnormal application counts, departmental moves, and various racial group representations in art units show right-skewed distributions, indicating a higher frequency of lower values. The `avg_woman_ratio` has a more symmetrical, bell-shaped distribution, suggesting a relatively even spread of female representation across different art units. Tenure days exhibit a long-tail distribution, implying that a smaller number of examiners have very long tenures. Variables representing the mean numbers of new, pending, and issued applications have different spread and skewness, which may correlate to the workflow patterns within the firm. The `examiner_id` variable shows a uniform spread, which is expected for a unique identifier. Overall, these distributions provide a foundational understanding of each variable's behavior for further multivariate analysis and modeling.

The bar charts (See Figure A.4) indicate frequency distributions for the categorical variables `gender`, `race`, and `start_year`, along with the `separation_indicator_sum`. The `gender` distribution shows a higher number of males compared to females. The `race` chart reveals that whites are the most represented race, followed by Asians, with other racial groups having significantly lower representation. The `start_year` variable indicates a concentration of starts in the early 2000s, with a peak around 2000, and fewer starts in more recent years. Lastly, the `separation_indicator_sum` suggests that the majority of the observations in the dataset are of examiners who have not separated (value 0), with fewer cases of separation (value 1). This data can help in assessing the demographic composition of the workforce and understanding employment trends over time within the organization.

2.3 Models

After performing the data preprocessing steps, we performed a survivorship analysis and drew the results in a Kaplan-Meier Survival Curve (See Figure A.5). Overall, we observed that the survivor curve is relatively smooth, but the employee drop off rate seems to accelerate as the tenure days increase.

For our main analysis, we opted for logistic regression models due to their efficacy in predicting binary outcomes like our current scenario. Our methodology involved the sequential development of four distinct models, each incorporating an additional layer of variables to progressively deepen our understanding of the factors influencing patent examination outcomes. Refer to Figure A.6) for all Model formulations.

Model 1 focused on basic demographic factors, including the gender and race of individual examiners. This foundational model aimed to assess the baseline influence of personal demographics on examination results.

Model 2 introduced job volume-related factors, such as the total number of applications processed and the number abandoned. This expansion allowed us to evaluate the impact of workload and productivity metrics alongside demographic characteristics.

Model 3, we broadened our scope to include department-wide variables, like the AU (Art Unit) mobility indicator and demographic ratios within the department (e.g., gender and race composition). This addition aimed to capture the broader organizational context and its potential effects on individual examiner outcomes.

Model 4 further advanced our analysis by integrating interactive variables between individual demographic characteristics and department-wide demographic details. This model was designed to uncover the nuanced interplay between personal and organizational demographics, offering insights into how individual

backgrounds might interact with the broader workplace environment to influence examination processes. Through this structured approach, these four models enable us to systematically dissect the complex dynamics at play in patent examination outcomes. Starting from individual characteristics and progressively incorporating job-related metrics and organizational contexts, we aim to reveal a multifaceted view of the factors that drive employee attrition patterns among patent examiners. Last but not least, we developed a random forest model that takes every factors in model 4 into account and graphed the feature importance scores in order to provide more support for the influence factors identified in the four models.

2.4 Evaluation and Results

2.4.1 ROC comparison

In evaluating the performance of the four predictive models, the Area Under the Receiver Operating Characteristic Curve (AUC) was used to measure the ability of each model to distinguish between the two classes of the dependent variable. See Figure A.7 The ROC curve is a plot of the true positive rate against the false positive rate at various threshold settings, and the AUC represents the degree to which the model is capable of correctly classifying outcomes.

See comparison in Table A.1

Model 1 and Model 2 both achieved an AUC of approximately 0.756, indicating a good level of discrimination that is significantly better than chance (which would have an AUC of 0.5). However, Model 3 and Model 4 demonstrated superior performance, with AUC values of 0.869 and 0.867, respectively. These higher AUC values suggest that Models 3 and 4 have a stronger discriminative ability compared to Models 1 and 2.

The ROC curves, previously discussed, visually represent these findings. The curves for Model 3 and Model 4 are closer to the top-left corner of the ROC space, reflecting their higher true positive rates and lower false positive rates across various decision thresholds. This visual representation complements the quantitative AUC values, offering a clear demonstration of Models 3 and 4's improved performance in predicting the outcome variable.

In summary, while all models show a capacity to differentiate between the outcomes better than random guessing, Models 3 and 4, with their higher AUC values, are more effective in this regard. Thus, they may be preferable for scenarios where accurately predicting the attrition indicator is critical.

2.4.2 Regression Summary

Refer to Figure A.8 and Figure A.9 for regression summary

Looking at the regression output, we noted the following key findings:

- **Tenure:** Longer tenure consistently showed a statistically significant negative association with the odds of attrition, indicating that each additional day of tenure slightly decreases the likelihood of attrition.
- **Start Year:** There was a strong and significant negative association between start year and the odds of attrition, with more recent start years being associated with higher odds of attrition.
- **Applications:** Variables related to applications (new, issued, and abnormal) introduced in Model 2 and beyond were significantly associated with the odds of attrition. Interestingly, issued applications were associated with an increase in attrition odds, while new and abnormal applications were associated with a decrease.
- **Art Unit and Movement:** Additional variables included in Model 3, such as the average number in the art unit and movement indicator, were significant, with the movement indicator negatively associated with attrition odds.

- **Interactions:** Model 4 revealed that the interaction of gender with the average minority ratio was significant, indicating that the relationship between gender and the attrition odds varied by the minority ratio within the group.
- **Gender:** Being male was associated with a slight increase in the odds of attrition, with the coefficient ranging from 0.052 to 0.637 across the models. However, this was not statistically significant in Models 1 through 3.
- **Race:** The coefficients for race suggest varying impacts on attrition odds. Notably, being Black was associated with a decrease in the odds of attrition compared to the reference category, though this was not statistically significant.

Model Performance

- Model complexity increased from Model 1 to Model 4, with log-likelihood improving, indicating a better fit with the addition of variables.
- The Akaike Information Criterion (AIC) decreased with each subsequent model, suggesting improved model quality with the inclusion of additional predictors.

Building on the logistic regression model summary, a Random Forest model was employed (See Figure A.10 to identify and quantify the importance of each feature in predicting attrition. The feature importance analysis revealed several key predictors with substantial influence on the attrition indicator:

- **PEN_applications_mean:** This feature had the highest importance score, indicating its significant role in predicting attrition. A high mean of pending applications seems to be a strong predictor of attrition.
- **new_applications_mean and ISSUED_applications_mean:** Both features also showed high importance, further suggesting that application-related metrics are critical in understanding attrition dynamics.
- **tenure_days:** Consistent with the logistic regression analysis, tenure days is identified as a significant predictor, reinforcing its negative association with the odds of attrition.
- **start_year:** The start year of employment, particularly recent years, was found to be highly predictive, aligning with the logistic regression findings where recent start years were associated with higher probability of attrition.

2.4.3 Conclusion

The analysis indicates that several factors are predictive of attrition. Work-related factors, especially tenure and start year, play a significant role in predicting attrition. Demographic factors showed varied influence, with some significant interaction effects. The evolving significance and magnitude of coefficients across the models suggest that the predictors have complex, interrelated impacts on the likelihood of attrition.

Further research should focus on understanding the causal mechanisms behind these associations and exploring potential strategies to reduce the likelihood of attrition, where desirable.

The feature importance from the Random Forest model corroborates the findings from the logistic regression analysis, emphasizing the relevance of tenure, start year, and application-related variables in predicting employee attrition. The convergence of evidence from these different analytical approaches strengthens the confidence in these predictors' roles. Organizational strategies to mitigate attrition may benefit from focusing on the application process and employee tenure. Interventions designed to address these areas could potentially lead to a reduction in employee attrition rates.

3 Part 2

In the realm of people analytics, there is a variety of tools that offer unique strengths tailored to this type of organizational needs. Among these, Crunchr emerges as a notable solution for its application of AI in uncovering associations and causal relationships between workforce dynamics and organizational objectives. This tool is designed to act as an AI-powered digital co-pilot, dubbed Crunchr Assistant, capable of dissecting workforce data across demographics, performance, and other relevant factors. This analytical prowess allows organizations to pinpoint the organizational and social causes of attrition, including workload, job satisfaction, engagement, and leadership styles. Furthermore, Crunchr's predictive analytics facilitate the anticipation of attrition risks, enabling preemptive measures. Benchmarking features also permit comparisons of attrition rates against industry standards, offering insights into whether observed trends are unique to the organization or indicative of broader industry phenomena.

Despite its advantages, the deployment of Crunchr is not without challenges. The accuracy and fairness of its analytics are contingent upon the quality of the underlying data. Biased or incomplete data sets risk entrenching inequalities and impairing decision-making processes. Additionally, the opacity of Crunchr's algorithmic operations (black-box) raises concerns about the justification of workforce decisions derived from its insights. There's also a danger in becoming overly dependent on technological solutions like Crunchr, potentially neglecting the invaluable role of human judgment in nuanced decision-making scenarios, particularly in complex environments such as the USPTO.

Recognizing these limitations, the team proposed an alternative solution, Patentlens, which offers a more targeted approach to addressing specific challenges faced by USPTO. Patentlens focuses on alleviating cognitive burdens associated with the volume of patent applications and ensuring consistency in evaluations. It accomplishes this by maintaining a curated, dynamically updated knowledge base that encompasses recent patents, legal rulings, and technological developments. This foundation enhances the AI's training and insights, ensuring that examiners are equipped with the most current and relevant information. More importantly, through presenting multiple aspects of the recommendation and providing detailed explanation, it removes the black-box nature of the algorithm and allows the user to make a more informed decision.

The juxtaposition of Crunchr and Patentlens in the context of people analytics underscores a critical consideration for organizations like the USPTO: the need to balance the benefits of AI-driven analytics with the imperatives of data quality, algorithmic transparency, and the irreplaceable value of human expertise. Furthermore, no matter which tool the department chose at the end, both highlight the importance of ethical considerations and privacy compliance in handling sensitive workforce data, emphasizing the need for a thoughtful and balanced approach to the adoption of people analytics solutions.

A Appendix

Figure A.1: Variables

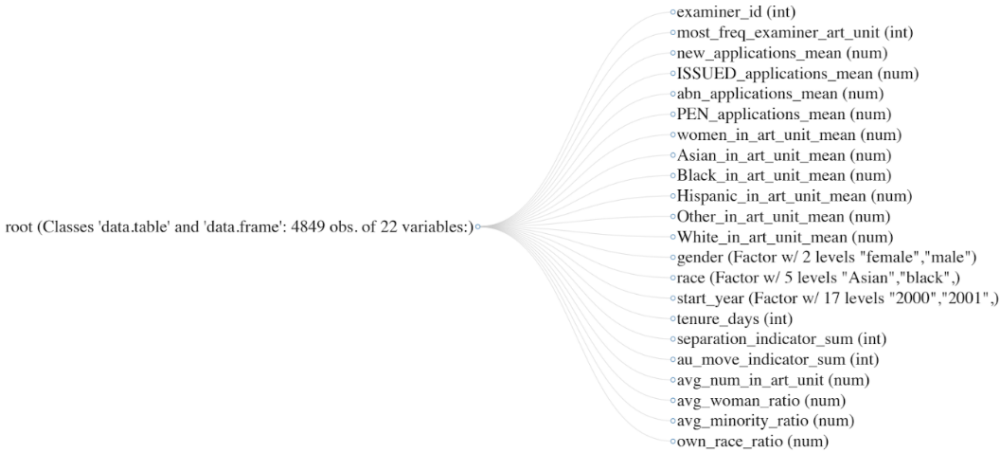

Figure A.2: Null counts

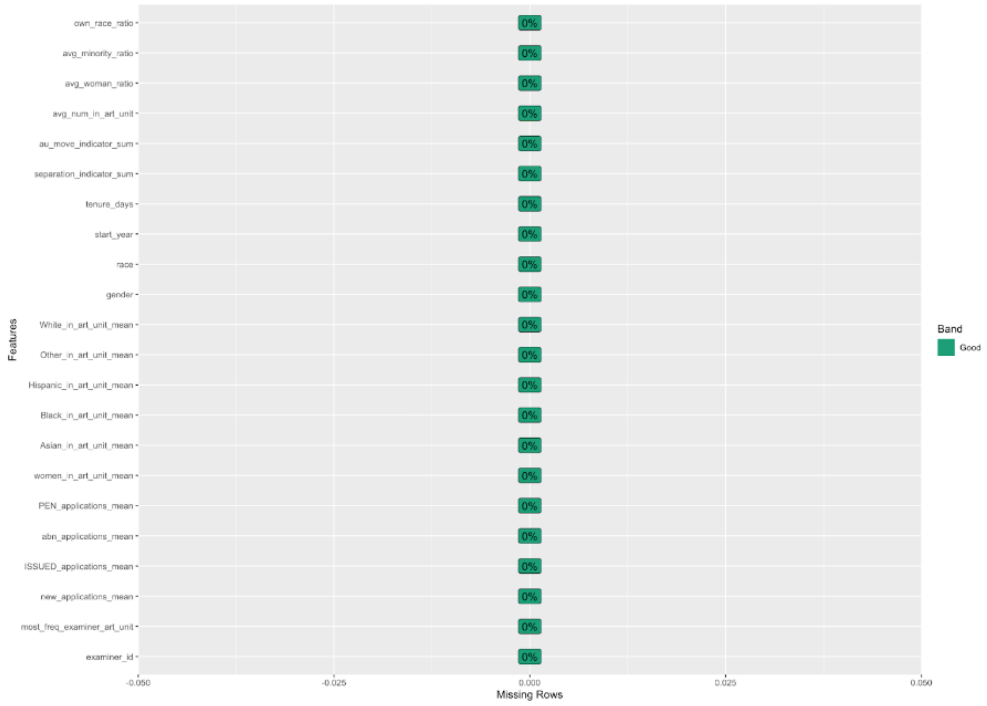

Figure A.3: Univariate Distribution

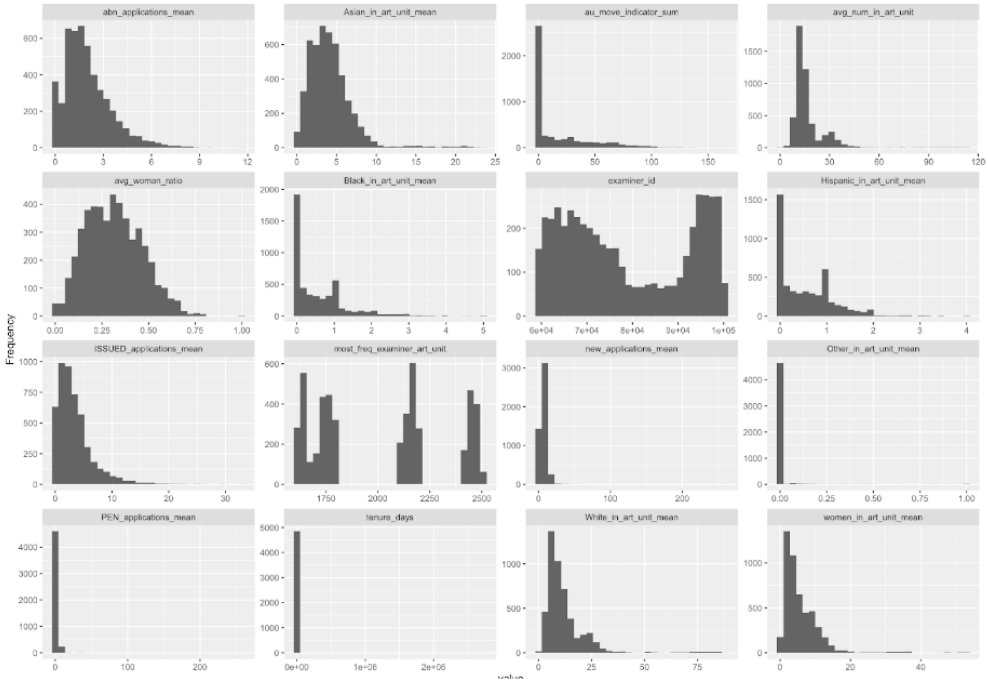

Figure A.4: Bar Frequency

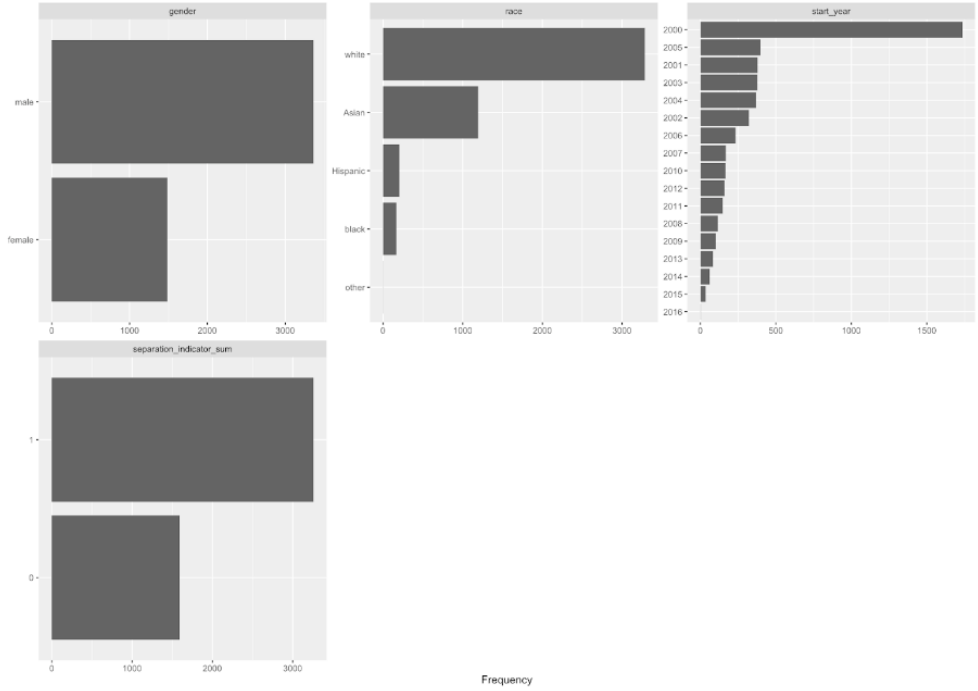

Figure A.5: Kaplan-Meier Survival

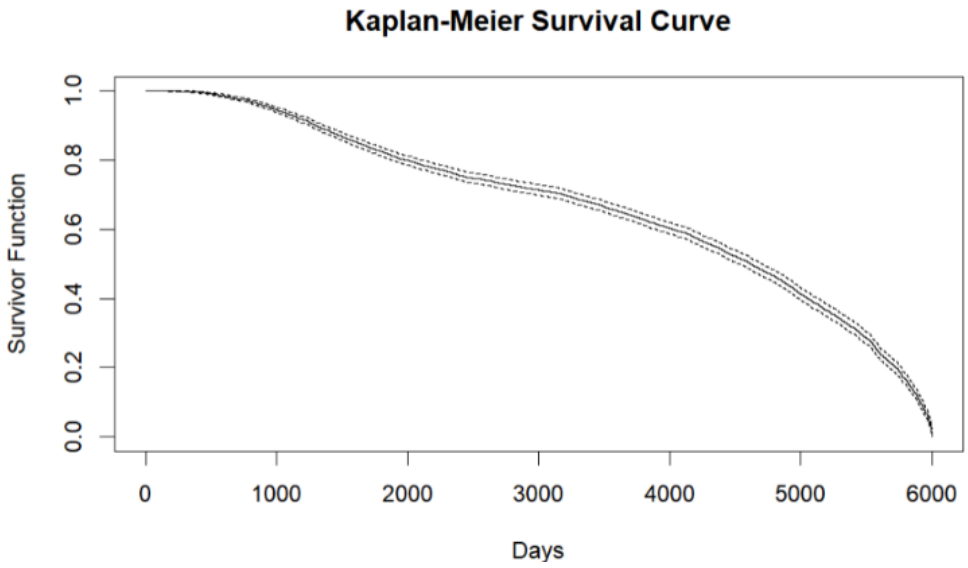

Figure A.6: Models

Model 1: Basic Demographics

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_{\text{gender}} + \beta_{\text{race}} + \beta_{\text{tenure}} + \beta_{\text{start}}$$

Model 2: Including Application Metrics

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \text{Model 1} + \beta_{\text{new\_apps}} + \beta_{\text{issued\_apps}} + \beta_{\text{abn\_apps}} + \beta_{\text{pen\_apps}}$$

Model 3: Adding Art Unit Information

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \text{Model 2} + \beta_{\text{au\_move}} + \beta_{\text{avg\_unit}} + \beta_{\text{woman\_ratio}} + \beta_{\text{minority\_ratio}} + \beta_{\text{own\_race}}$$

Model 4: Including Interactions

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \text{Model 3} + \beta_{\text{gender}\times\text{woman\_ratio}} + \beta_{\text{gender}\times\text{minority\_ratio}}$$

Table A.1: AUC scores for different models.

Model	AUC
Model 1	0.75570256556172
Model 2	0.75570256556172
Model 3	0.868549343197229
Model 4	0.866861099255464

Figure A.7: Models

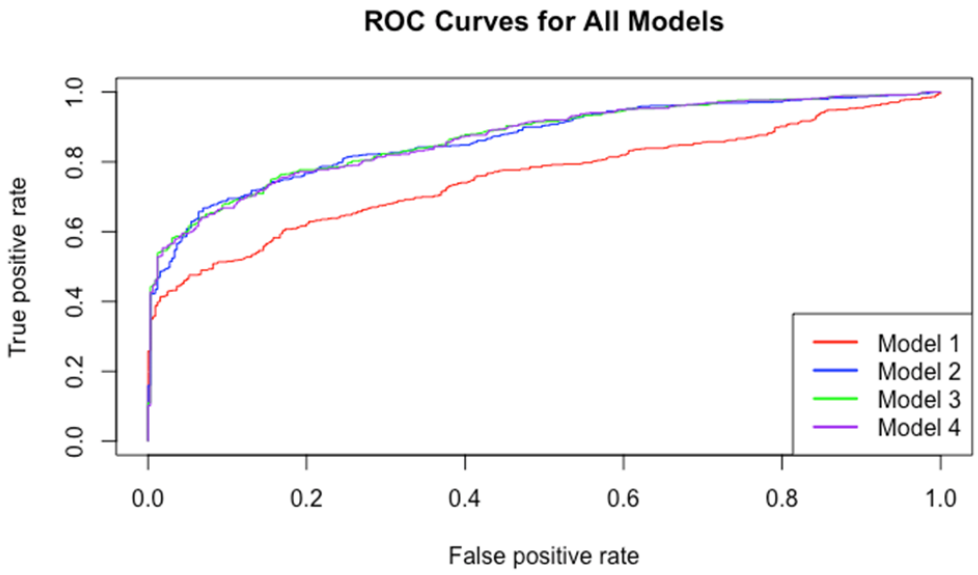

Figure A.8: Model Summary

Comparative Logistic Regression Model Summary

	Dependent variable:			
	separation_indicator_sum			
	(1)	(2)	(3)	(4)
gendermale	0.130	0.096	0.052	0.637
	(0.080)	(0.086)	(0.096)	(0.403)
raceblack	-0.100	-0.012	-0.114	-0.133
	(0.215)	(0.228)	(0.245)	(0.245)
raceHispanic	0.149	0.180	0.053	0.044
	(0.203)	(0.216)	(0.236)	(0.236)
raceother	0.068	0.084	-0.022	0.002
	(1.431)	(1.460)	(1.516)	(1.517)
racewhite	0.067	0.176*	0.123	0.121
	(0.086)	(0.093)	(0.140)	(0.140)
tenure_days	-0.003***	-0.002***	-0.001***	-0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
start_year2001	-1.665***	-1.116***	-0.982***	-0.986***
	(0.178)	(0.188)	(0.187)	(0.187)
start_year2002	-2.929***	-1.881***	-1.554***	-1.558***
	(0.236)	(0.246)	(0.238)	(0.238)
start_year2003	-3.968***	-2.406***	-1.807***	-1.802***
	(0.298)	(0.314)	(0.293)	(0.293)
start_year2004	-5.145***	-2.987***	-2.287***	-2.287***
	(0.367)	(0.388)	(0.358)	(0.358)
start_year2005	-5.926***	-3.171***	-2.198***	-2.205***
	(0.442)	(0.469)	(0.429)	(0.430)
start_year2006	-6.944***	-3.481***	-2.276***	-2.269***
	(0.527)	(0.561)	(0.512)	(0.512)
start_year2007	-7.569***	-3.590***	-2.166***	-2.171***
	(0.603)	(0.648)	(0.594)	(0.594)
start_year2008	-8.633***	-3.912***	-2.202***	-2.188***
	(0.690)	(0.745)	(0.683)	(0.683)
start_year2009	-10.219***	-4.958***	-2.914***	-2.922***
	(0.774)	(0.828)	(0.757)	(0.758)
start_year2010	-11.010***	-5.285***	-2.968***	-2.960***
	(0.835)	(0.893)	(0.814)	(0.815)
start_year2011	-11.522***	-5.005***	-2.464***	-2.461***
	(0.906)	(0.982)	(0.897)	(0.898)
start_year2012	-12.897***	-5.648***	-2.876***	-2.875***
	(0.985)	(1.065)	(0.972)	(0.972)

Figure A.9: Model Summary Cont.

start_year2013	-13.752***	-5.281***	-2.202**	-2.216**
	(1.085)	(1.195)	(1.102)	(1.103)
start_year2014	-12.827***	-2.262	1.223	1.223
	(1.327)	(1.508)	(1.437)	(1.437)
start_year2015	-14.473***	-2.438	1.524	1.542
	(1.394)	(1.849)	(1.894)	(1.904)
start_year2016	-4.362	5.200	10.219	10.264
	(378.372)	(183.784)	(303.356)	(303.376)
new_applications_mean		-0.749***	-0.815***	-0.816***
		(0.049)	(0.052)	(0.052)
ISSUED_applications_mean		0.666***	0.689***	0.690***
		(0.049)	(0.051)	(0.051)
abn_applications_mean		0.684***	0.642***	0.647***
		(0.067)	(0.069)	(0.069)
PEN_applications_mean				
au_move_indicator_sum			-0.024***	-0.024***
			(0.002)	(0.002)
avg_num_in_art_unit			0.028***	0.029***
			(0.006)	(0.007)
avg_woman_ratio			-0.598	-0.192
			(0.370)	(0.618)
avg_minority_ratio			-1.078***	-0.248
			(0.352)	(0.571)
own_race_ratio			-0.237	-0.226
			(0.314)	(0.314)
gendermale:avg_woman_ratio				-0.609
				(0.736)
gendermale:avg_minority_ratio				-1.137*
				(0.615)
Constant	18.955***	12.760***	9.455***	9.029***
	(1.341)	(1.394)	(1.278)	(1.311)
Observations	3,880	3,880	3,880	3,880
Log Likelihood	-2,074.028	-1,872.100	-1,776.118	-1,774.400
Akaike Inf. Crit.	4,194.057	3,796.200	3,614.237	3,614.800
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure A.10: Model Summary Cont.