Final Group Project

Data analysis of the US.Patent and Trademark Office

Tarek Cheaito

Rohana Habib

Himanshu Mayank

Xénia Sozonoff

2023-04-18

\

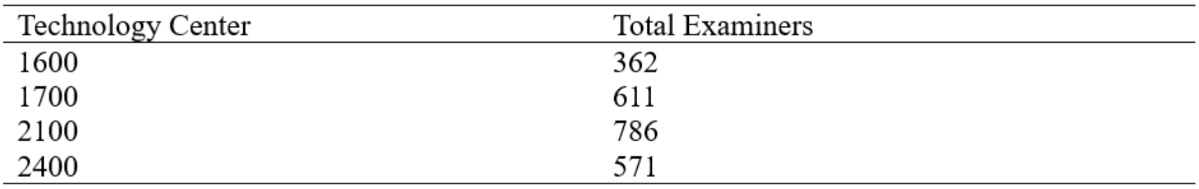
Introduction

Through this project we will thoroughly analyze patent application data and examiner information from various public sources. We will focus on investigating the factors that influence the duration of patent application prosecution, paying particular attention to the impact of network structure, race, and ethnicity. Specifically, we will be examining a subset of patent applications filed after 2000 in four technology centers of the agency. We hope to uncover potential disparities in the examination process, make recommendations to reduce the backlog, and ensure a fair and efficient process. Our project seeks to promote a more equitable and fair system that supports economic growth and encourages innovation.

In addition to traditional statistical analysis, we will utilize organizational network analysis (ONA) techniques to gain a deeper understanding of the relationships and interactions among examiners and patent applications. By examining the network structure and applying centrality measures, we can identify key players and bottlenecks that may be contributing to longer processing times. This approach will enable us to gain insights into the examination process that may not be readily apparent through traditional methods. Through this interdisciplinary approach, we hope to provide a more comprehensive understanding of the factors that impact patent application processing times and inform policy recommendations to improve the system.

Exploratory Analysis

Figure 1: Number of Examiners by Technology Center



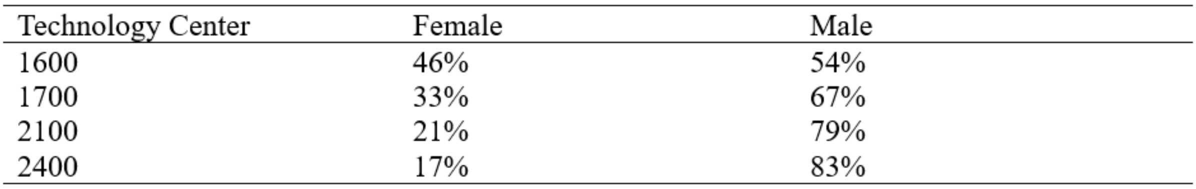
Technology Center	Total Examiners
1600	362
1700	611
2100	786
2400	571

Note: Individual art units are split into four technology centers based on what patent applications they examine:

Center 1600: Biotechnology & Organic fields
Center 1700: Chemical & Materials Engineering fields
Center 2100: Computer Architecture Software & Information Security
Center 2400: Computer Networks, Multiplex, Cable & Cryptography/Security

During the exploratory phase, we present percentages rather than counts when comparing characteristics such as gender, race, and tenure days among examiners. As shown in Figure 1 the number of examiners is not proportional across technology centers, which could result in misleading comparisons if presented solely based on counts. By using percentages, we can make more accurate comparisons that take into account the varying examiner populations across technology centers.

Figure 2: Examiners by Gender

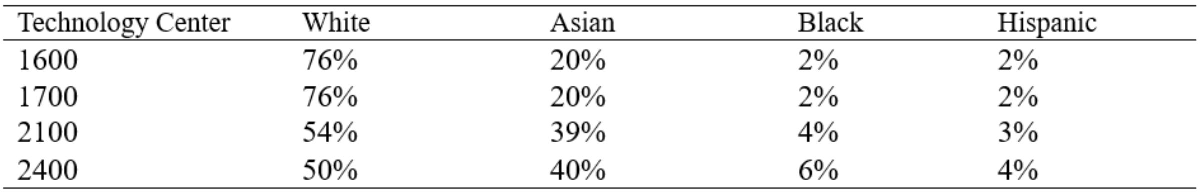


Technology Center	Female	Male
1600	46%	54%
1700	33%	67%
2100	21%	79%
2400	17%	83%

Note: For the purpose of clarity, percentages presented in this report have been rounded to the nearest whole number. However, this rounding does not affect the overall trends and patterns identified in the data.

Among all the technology centers, the number of male examiners exceeds that of female examiners. However, when examining the breakdown of male and female examiners across individual centers, we found notable disparities. In particular, centers 1700, 2100, and 2400 had exceptionally low female representation. Center 1600, on the other hand, had a more balanced gender ratio, with 45.58% of examiners being male and 54.42% being female.

Figure 3: Examiners by Race

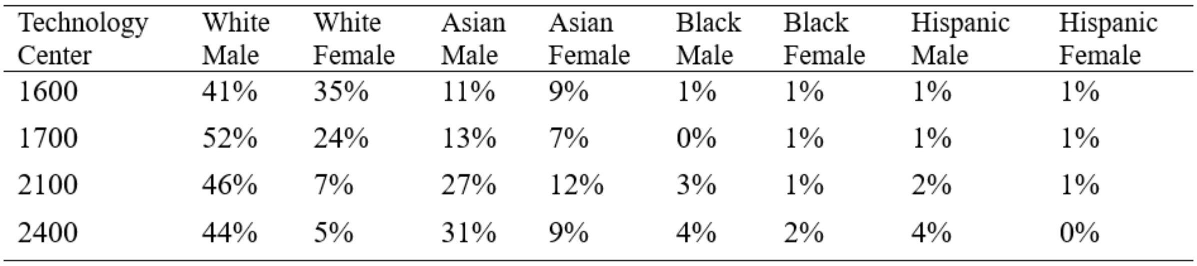


Technology Center	White	Asian	Black	Hispanic
1600	76%	20%	2%	2%
1700	76%	20%	2%	2%
2100	54%	39%	4%	3%
2400	50%	40%	6%	4%

Note: The 'Other' race category has been removed from this figure as these individuals accounted for less than 1% of the population

Although technology center 1600 had a relatively balanced ratio of female to male examiners, it shares a similar pattern with center 1700 in having a predominantly White examiner population, with little representation from Black or Hispanic groups. To note, center 1600 and 1700 have approximately the same race breakdown. Meanwhile, for centers 2100 and 2400, around half of their examiners are White, 40% are Asian, and the remainder are distributed between Black and Hispanic groups.

Figure 4: Examiners by Race & Gender

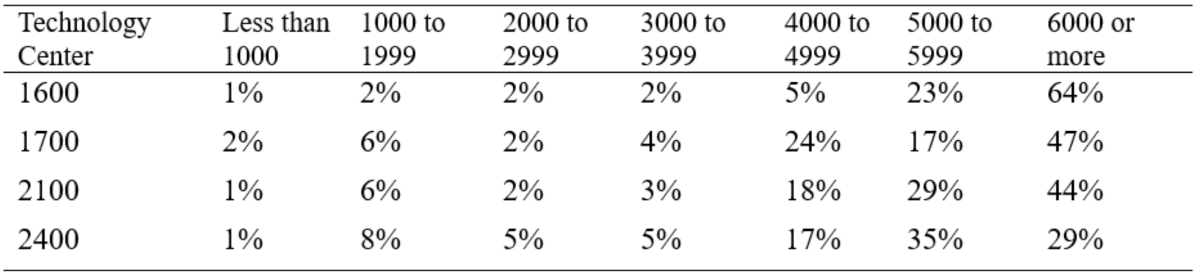


Technology Center	White Male	White Female	Asian Male	Asian Female	Black Male	Black Female	Hispanic Male	Hispanic Female
1600	41%	35%	11%	9%	1%	1%	1%	1%
1700	52%	24%	13%	7%	0%	1%	1%	1%
2100	46%	7%	27%	12%	3%	1%	2%	1%
2400	44%	5%	31%	9%	4%	2%	4%	0%

Note: As previously stated, percentages have been rounded for clarity. Therefore, the total percentage might not always add up to exactly 100%. This rounding does not affect the overall trends identified in the data.

Figure 4 provides a view of the distribution of examiners across gender (Figure 2) and race (Figure 3) categories. Pairing up race and gender in one table offers a more granular perspective on the intersectionality of race and gender in each technology center.

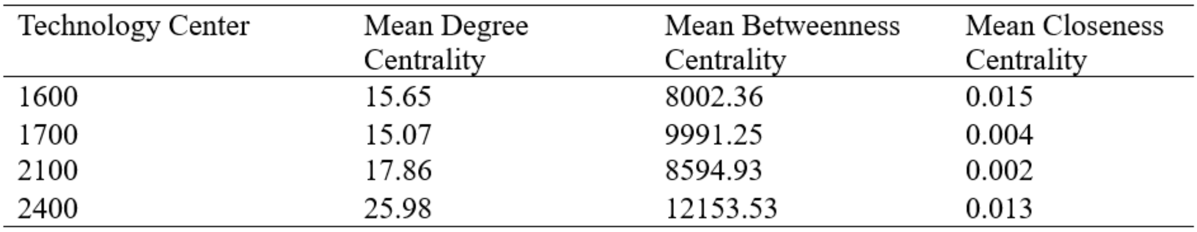
Figure 5: Examiners by Grouped Tenure Days



Technology Center	Less than 1000	1000 to 1999	2000 to 2999	3000 to 3999	4000 to 4999	5000 to 5999	6000 or more
1600	1%	2%	2%	2%	5%	23%	64%
1700	2%	6%	2%	4%	24%	17%	47%
2100	1%	6%	2%	3%	18%	29%	44%
2400	1%	8%	5%	5%	17%	35%	29%

The data indicates that most examiners across all technology centers have been with the USPTO for 5000 or more tenure days. This trend suggests that most examiners have chosen to remain with the agency for a considerable period, which could speak to the quality of the working environment or the opportunities for career growth and professional development that the USPTO offers.

Figure 6: Average Centralities

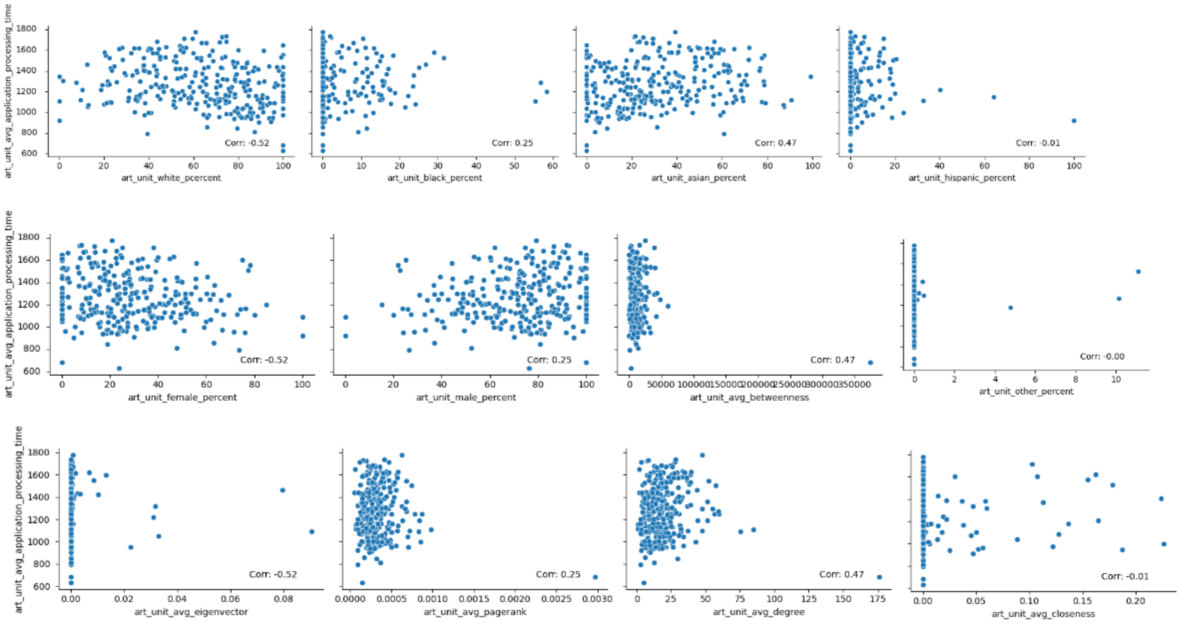


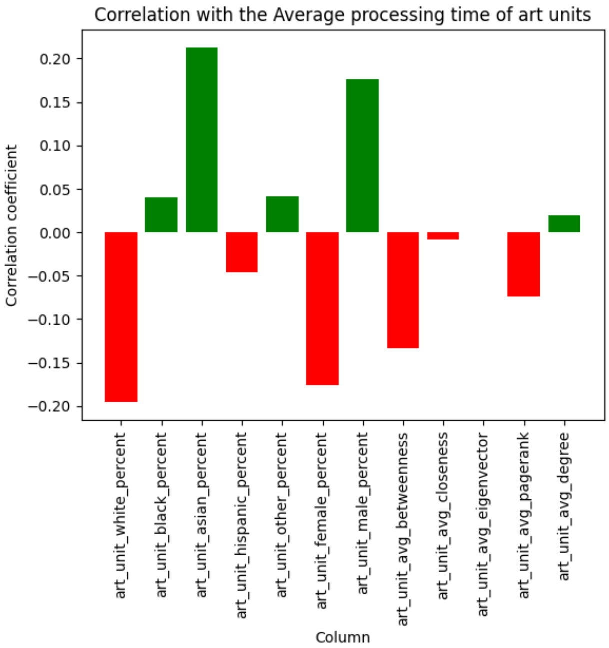
Technology Center	Mean Degree Centrality	Mean Betweenness Centrality	Mean Closeness Centrality
1600	15.65	8002.36	0.015
1700	15.07	9991.25	0.004
2100	17.86	8594.93	0.002
2400	25.98	12153.53	0.013

Based on the centralities of the examiners in the network, we can draw some conclusions about the network structure. Specifically, examiners in technology centers 2100 and 2400 have relatively high degree centralities, with center 2400 having the highest mean degree centrality of around 25.98. This suggests that these examiners have the most connections with other examiners in the network. Examiners in center 2400 have the highest betweenness centralities, suggesting that they play an important role as bridges between other pairs of examiners in this network. Examiners in technology centers 1600 and 2400 have relatively high closeness centrality, indicating that they are relatively close to other examiners in the network, while examiners in groups 1700 and 2100 have a very low closeness centrality, indicating that they are very far away from other examiners in the network.

Modeling & Insights

Figure 7: Correlation of the variables with the application processing time



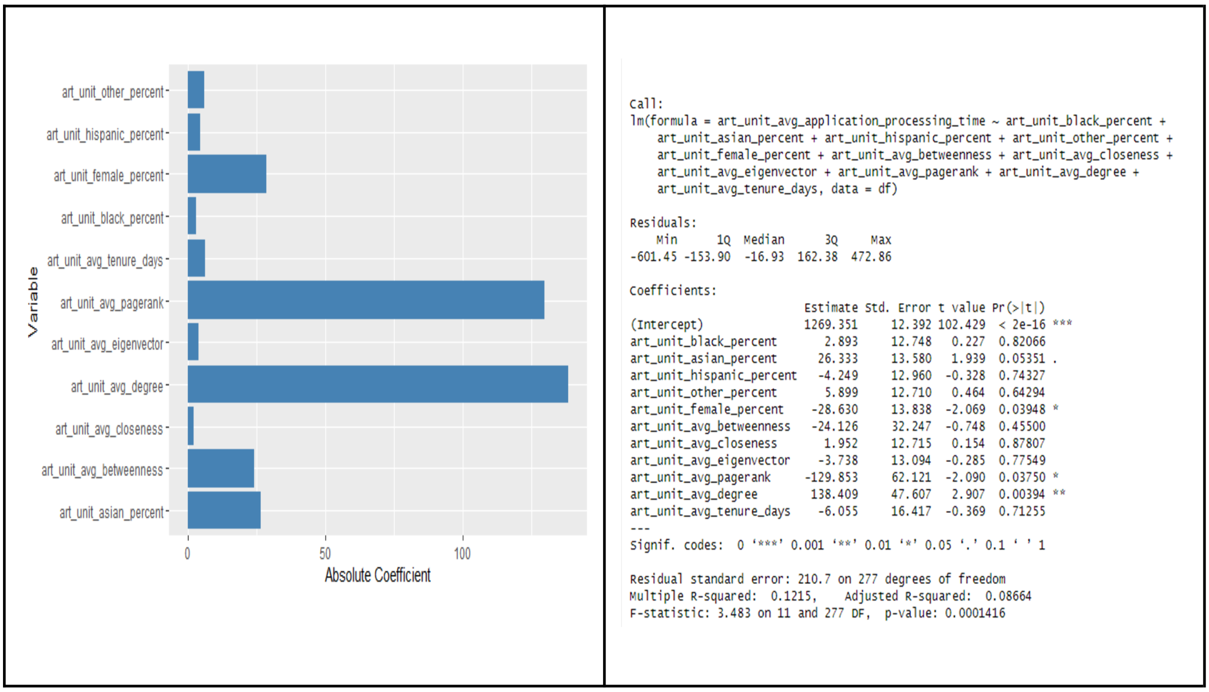


The average Asian percentage, black percentage, other race percentages, average degree percentage, and average male percentage positively correlate with the application processing time. The average white percentage, Hispanic percentage, female percentage, closeness centrality, betweenness centrality and PageRank centrality have negative correlation with the application processing time.

Linear Regression

After conducting an exploratory analysis of the patent application and examiner data, we then used logistic regression to model the relationship between the independent variables and art_unit_avg_application_processing_time. Our goal is to identify which factors are statistically significant predictors of processing times and gain further insights into potential inequities within the examination process, enabling us to gather insight into the factors that have a positive or negative impact on the average application processing time.

Figure 7: Linear Regression and Variable Importance Plot

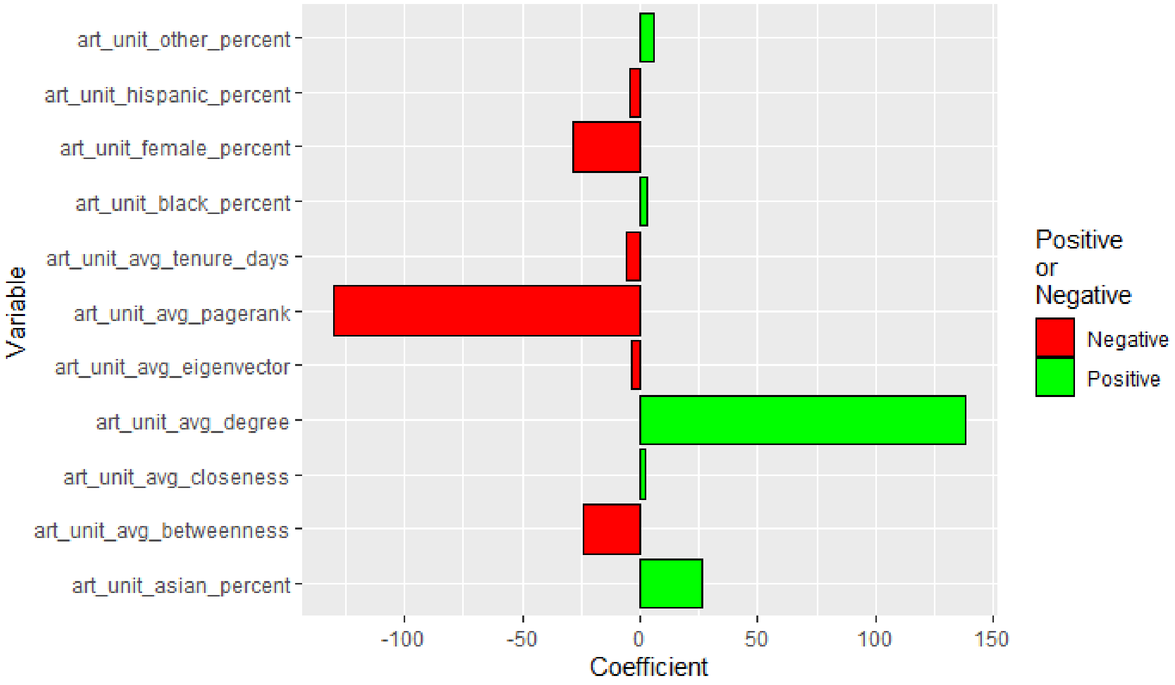


```
Call:
lm(formula = art_unit_avg_application_processing_time ~ art_unit_black_percent +
    art_unit_asian_percent + art_unit_hispanic_percent + art_unit_other_percent +
    art_unit_female_percent + art_unit_avg_betweenness + art_unit_avg_closeness +
    art_unit_avg_eigenvector + art_unit_avg_pagerank + art_unit_avg_degree +
    art_unit_avg_tenure_days, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-601.45 -153.90  -16.93  162.38  472.86

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1269.351     12.392 102.429  < 2e-16 ***
art_unit_black_percent       2.893     12.748   0.227  0.82066
art_unit_asian_percent      26.333     13.580   1.939  0.05351 .
art_unit_hispanic_percent   -4.249     12.960  -0.328  0.74327
art_unit_other_percent       5.899     12.710   0.464  0.64294
art_unit_female_percent    -28.630     13.838  -2.069  0.03948 *
art_unit_avg_betweenness   -24.126     32.247  -0.748  0.45500
art_unit_avg_closeness       1.952     12.715   0.154  0.87807
art_unit_avg_eigenvector    -3.738     13.094  -0.285  0.77549
art_unit_avg_pagerank     -129.853     62.121  -2.090  0.03750 *
art_unit_avg_degree        138.409     47.607   2.907  0.00394 **
art_unit_avg_tenure_days    -6.055     16.417  -0.369  0.71255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 210.7 on 277 degrees of freedom
Multiple R-squared:  0.1215,    Adjusted R-squared:  0.08664
F-statistic: 3.483 on 11 and 277 DF,  p-value: 0.0001416
```

The adjusted R-squared value (0.12) indicates that the model is not a very good fit for the data as it is small. It also articulates that the independent variables in the model explain only a small amount of the variation in the dependent variable. The average degree centrality is the one of the most statistically significant predictors of the dependent variable.The average female percentage within the art unit and the average pagerank centrality within the art unit are also some of the important variables for prediction.

To conclude, while this model provides some insights into the relationship between the independent and dependent variables, its relatively low explanatory power suggests that additional factors beyond those included in the model may also be important in predicting art_unit_avg_application_processing_time.

Figure 8: Effect of Variables on Application Processing Time



Based on the important variables and their effect on an art unit's average application processing time, the art units with a higher degree centrality have higher application processing time whereas the art unit with higher average page rank centrality has lower processing time. There could be various possible speculations for this.

- Art units with higher centrality may receive more applications, leading to longer processing times due to the increased workload. Individuals with higher degree of centrality may be more likely to receive a larger number of inquiries and requests for information from other network members, which could cause delays in processing applications.
  Individuals with higher degree centrality may have more responsibilities and require specialized training and experience, which could lead to longer processing times.
- Art units with higher centrality may be responsible for handling more complex or challenging applications, which could require more time and resources to process. This could be because these units have developed specialized expertise in certain technical areas or have a reputation for being more rigorous in their examination process.
- Art Units wherein the average value of the betweenness centrality is higher, the average processing time is lower; ie, the applications get processed faster.
  One possible reason for this is that units with higher betweenness centrality may have more efficient processes and communication channels, allowing for faster and smoother processing of patent applications. These units may also be more experienced and knowledgeable, which could contribute to faster processing times. Additionally, higher betweenness centrality may indicate that a unit is more connected to other units

within the patent examination network, which could facilitate collaboration and information sharing, leading to improved efficiency and faster processing times.

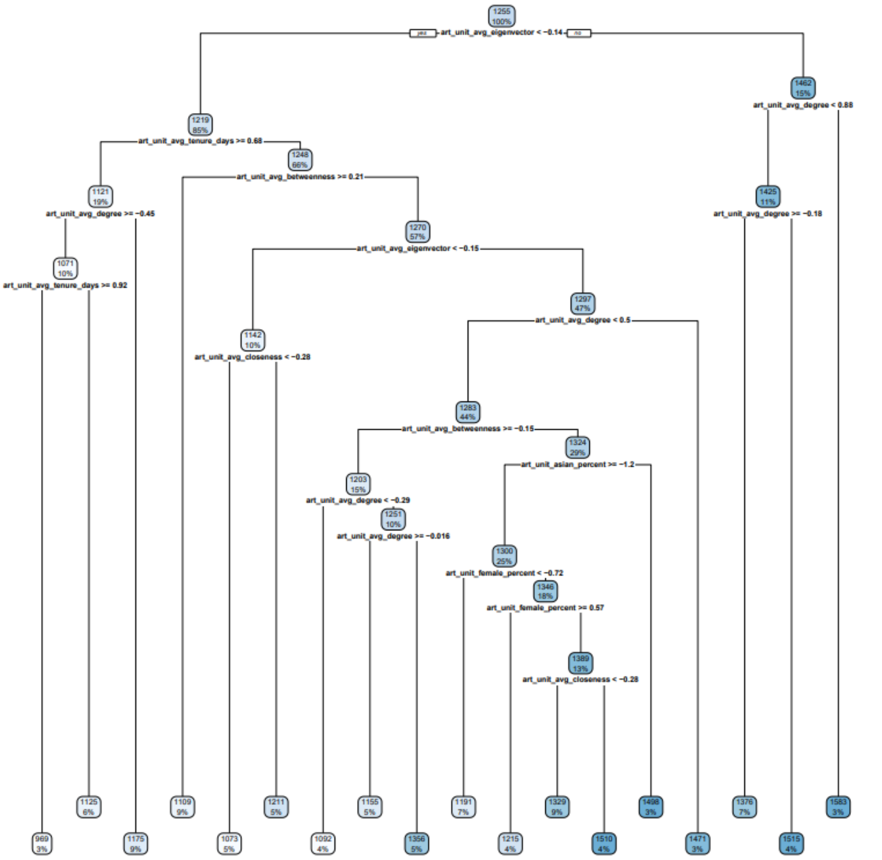
- Other factors contribute to the relationship between PageRank centrality and application processing time, such as differences in staffing levels, budget allocations, or administrative processes.

  Pagerank centrality is a measure of influence and importance in a network, so individuals with high PageRank centrality are likely to have access to more resources, information, and support from other influential individuals in the network. This may enable them to work more efficiently and quickly complete their tasks. Additionally, individuals with high PageRank centrality may have developed strong relationships and collaborations within the network, facilitating their work and leading to faster completion times.

Decision Tree and Random Forest

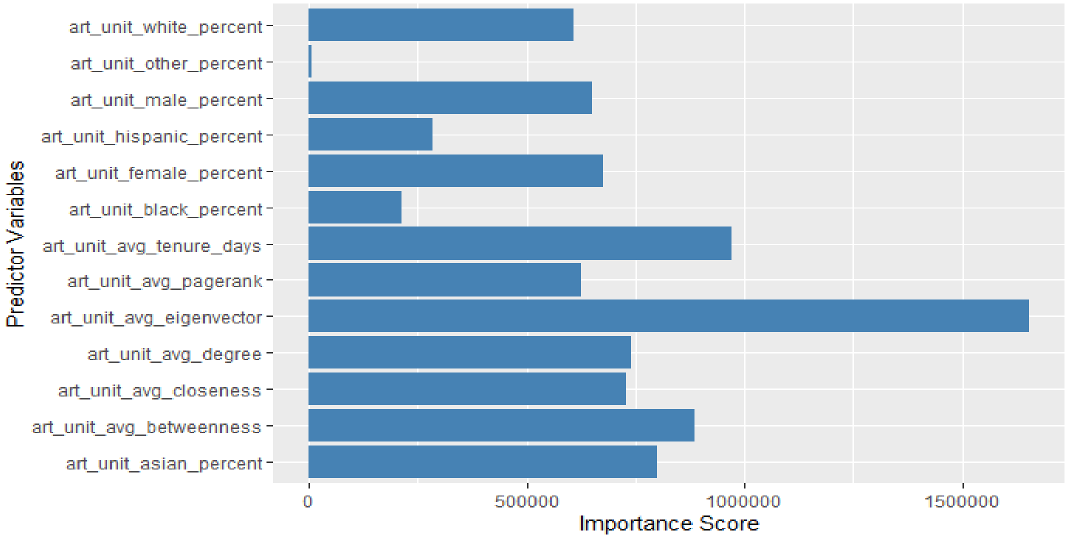
In addition to the logistic regression, we are also doing a Decision Tree and random Forest model to gain additional insights into the factors influencing patent application processing time. Unlike logistic regression, Decision Trees and Random forests can handle non-linear relationships and variable interactions. This model helps us to identify which variables are most important in predicting art_unit_avg_application_processing_time.

Figure 9: Decision Tree


1255
100%
yes
art_unit_avg_eigenvector < −0.14
no
1219
85%
art_unit_avg_tenure_days >= 0.68
1248
66%
art_unit_avg_betweenness >= 0.21
1121
19%
art_unit_avg_degree >= −0.45
1071
10%
art_unit_avg_tenure_days >= 0.92
1270
57%
art_unit_avg_eigenvector < −0.15
1142
10%
art_unit_avg_closeness < −0.28
1297
47%
art_unit_avg_degree < 0.5
1283
44%
art_unit_avg_betweenness >= −0.15
1324
29%
art_unit_asian_percent >= −1.2
1203
15%
art_unit_avg_degree < −0.29
1251
10%
art_unit_avg_degree >= −0.016
1300
25%
art_unit_female_percent < −0.72
1346
18%
art_unit_female_percent >= 0.57
1389
13%
art_unit_avg_closeness < −0.28
1462
15%
art_unit_avg_degree < 0.88
1425
11%
art_unit_avg_degree >= −0.18
969
3%
1125
6%
1175
9%
1109
9%
1073
5%
1211
5%
1092
4%
1155
5%
1356
5%
1191
7%
1215
4%
1329
9%
1510
4%
1498
3%
1471
3%
1376
7%
1515
4%
1583
3%

The average eigenvector centrality of the examiners, average tenure days, and average degree centrality within the art unit are the most important variables based on which the decision tree is split.

Figure 10: Variable Importance Plot



The average eigenvector centrality, average tenure days, and average betweenness centrality are the most important variables in the prediction task.

Recommendations & Value Proposition

Network analysis is a powerful tool for people analytics as it provides a window into the social structures and relationships within an organization. Examining network structure can help identify key individuals who significantly impact the organization's performance. Our network analysis of the USPTO provides actionable insights that can help the USPTO optimize application processing times and underscores the importance of the network structure in determining organizational outcomes.

To streamline application processing, the USPTO should leverage individuals with high betweenness centrality, who act as bridges in the network as they have proved to be effective at connecting and collaborating with others to produce more efficient outcomes. At the same time, the USPTO should monitor the impact of individuals with high degree centrality, who have many connections as they may slow down processing times. Moreover, building and maintaining a diverse workforce is crucial for improving processing times. Departments with a higher percentage of female individuals have faster processing times, highlighting the importance of prioritizing diversity and inclusivity. A diverse workforce brings unique perspectives and experiences that can lead to more efficient and effective application processing.

Furthermore, as high pagerank centrality and high degree centrality are extremely important in measuring application process time where pagerank centrality lowers application process time and degree centrality increases application process time the USPTO should attempt to convert those with high degree centrality into PageRank centrality. To achieve this the USPTO can foster strong connections between individuals with high degree centrality and those with high pagerank centrality. This can be done through encouraging individuals with high degree centrality to collaborate with those who have high PageRank centrality, which can increase the former's influence and improve their position in the network. Additionally, they can provide training and resources to help individuals with high degree centrality develop their leadership and communication skills, which may also increase their pagerank centrality. Bottlenecks and inefficiencies in the network causing high degree centrality individuals to negatively impact processing times should also be addressed. By remediating these bottlenecks and improving the flow of communication and collaboration, individuals with high degree centrality can become more effective in their roles, increasing their pagerank centrality and overall efficiency in terms of application process time. It is also good to note that individuals with higher pagerank centrality may have the ability to facilitate the spread of complex contagion in a network and actually produce change in behaviors.

Conclusion

Our analysis has provided valuable insights into the factors that impact the length of patent application prosecution and the potential biases within the examination process. Our model can help identify which technology centers are lagging in application processing. This information can aid in optimizing those centers by placing individuals with higher betweenness centrality or prioritizing hiring more female individuals. Identifying factors that lead to faster processing times will optimize resource allocation and ensure that valuable resources are allocated to the most important areas. Furthermore, identifying potential biases within the organization's processes will enable proactive measures to be taken to tackle discrimination and increase efficiency.