ORGB 671: Assignment 5
US Patent Office

Group 4:

Will Stephenson	261082056
Hugo Garcia	260791363
Liliana Tretyakova	261086215
Emery Dittmer	260658030

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 are examining the human metrics of these patent employees. This includes turnover rate, mobility within the company and more. To accomplish this we are using a dataset provided by the USPTO (simplified by our instructor). Additional details can be found here and the datasets here.

2.1. Dataset & Pre-Processing

We pre-processed the dataset to include all the patent transactions, patent examiner ID, gender and ethnicity and removed the application status dates that were missing as we are relying on this field for forecast predictions.

Next, since we are seeking to determine the production (number of application decisions by art unit in a week), we filtered the status update to reflect a decision. In the dataset there are 3 possible status updates: 'ISS' = Patent Issued, 'ABN' = Patent Abandoned, 'PEND' = Patent Pending. With this, we removed all the applications with a patent pending status, and this left us with 83.8% of our data remaining. We believe this is an acceptable amount for our analysis.

Next, we grouped our data based on processed applications, making a gender neutral and 2 gender inclusive (Male, Female) data sets.

2.2. Using the 'Prophet' Package to Predict

The input to Prophet is a dataframe with two columns: ds and y. The ds (datestamp) column is expected as YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric and represents the measurement we wish to forecast.

With our pre-processed datasets we created an empty dataframe called future which has 365 days and we used the prophet package to predict the rates and future state. With this, we predicted between 3,600 and 4,300 patent applications per week at the end of 2017, however, this is not accounting for holidays.

Based on these factors we achieved the following graph.

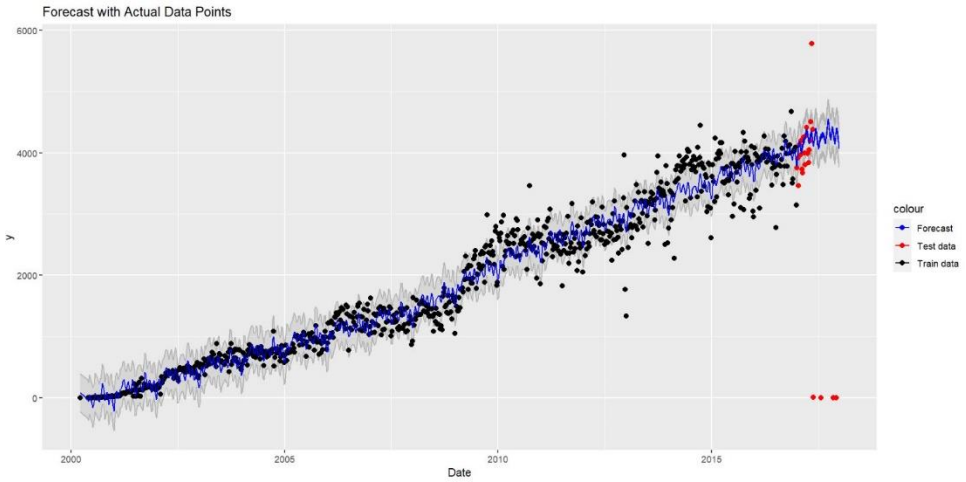

Figure 1: Prediction of 2017 Patent examiner production based on 2000 - 2016 historical Data.

The following statical measures evaluate the model and how accurate it is. This is best understood in the context of similar models. In general, a lower mean squared error (MSE) indicates better model performance.

Measure	*Value*	*Measure*	*Value*
Mean Squared Error (MSE)	*2,926,939*	*Root Mean Square Error (RMSE)*	*1,710.8*
Mean Absolute Error (MAE)	*980.99*	*Mean Absolute Percentage Error (MAPE)*	*516*

2.3. Analysis

Based on the graph, there are several concerning trends:

First, 2017 presents several outliers, as indicated in the figure below. We will need to eliminate these outliers to improve model performance.

Second, after evaluating the number of data points in 2017 there remained fewer than 52 (one for each week) in 2017. This likely indicates that these status updates were removed due to pre-processing or that the dataset itself does not have a complete data picture.

Lastly, a deeper dive into the data demonstrates numerous points during peak holidays and weekend times. Therefore, we will need to adjust the model assumptions and date methodology for the model to produce more accurate results.

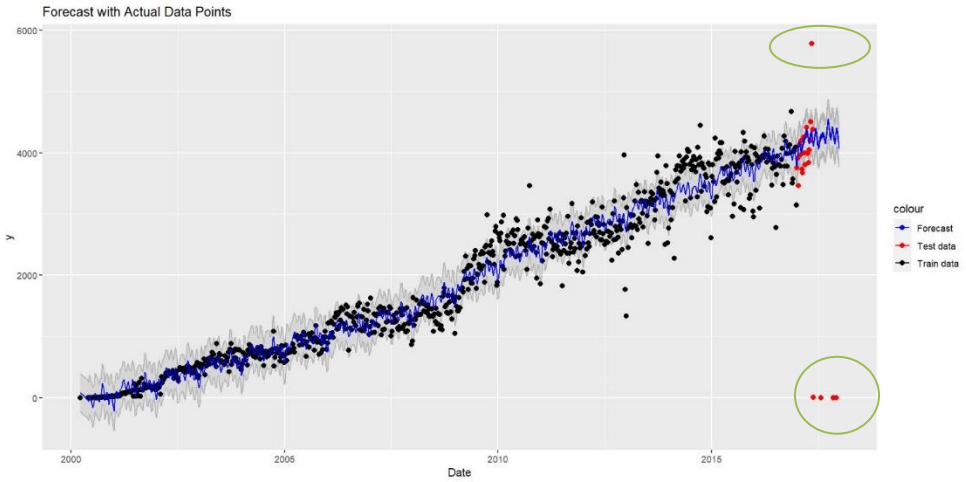

Figure 2: Prediction of 2017 Patent examiner production based on 2000 - 2016 historical Data. Outliers shown in green.

3.1. Accounting for Holidays

3.1.1. Background

Before making any drastic data manipulations to the data, we first attempted to improve the model's predictions by adding US public holidays. While this had an effect, it is clear, based on the model values, other tactics were still required.

We used the following two publicly available sources for adding US Public Holidays:

1. Kaggle Dataset of US Holiday Dates (2004 – 2021). Link here
2. US Department of Commerce – Federal Holidays. Link here

3.1.2. Plots

After predicting the trends, we plotted the values. The following plots are the result

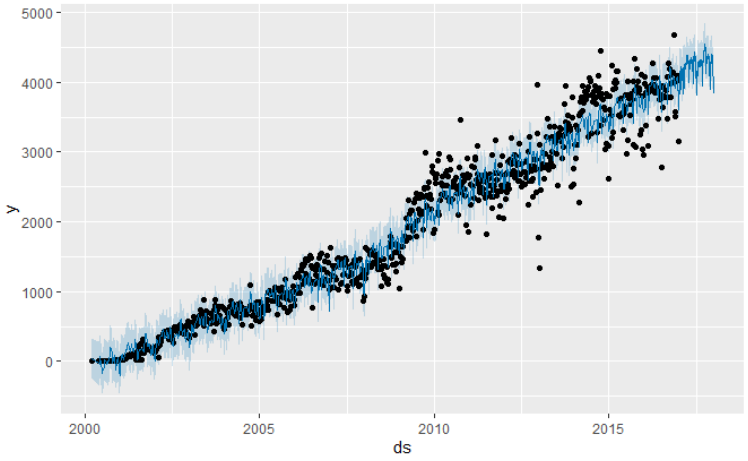

Figure 3: Prediction of 2017 Patent examiner production based on 2000 - 2016 historical Data and Accounting for Holidays.

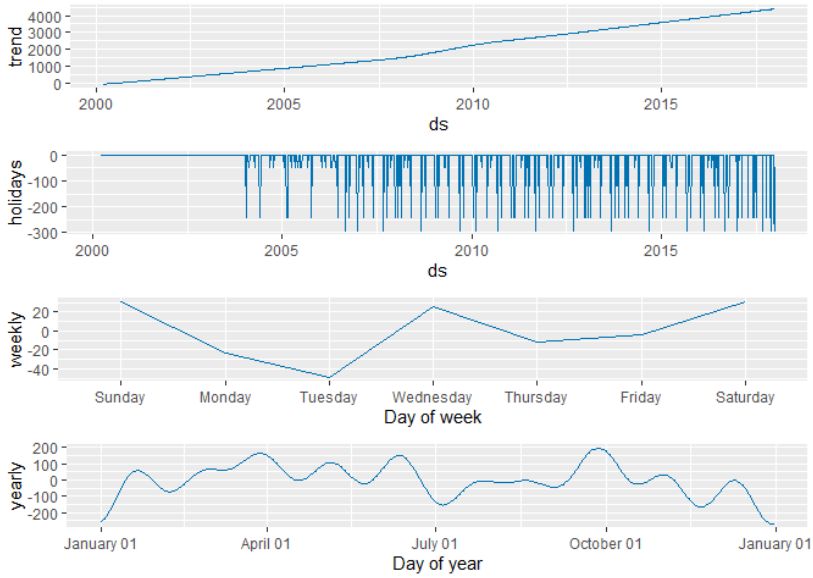

Figure 4: Component Plot of 2017 Holliday Predictions.
This Decomposes the data into Annual, Monthly, Day of Week and Holiday Trends.

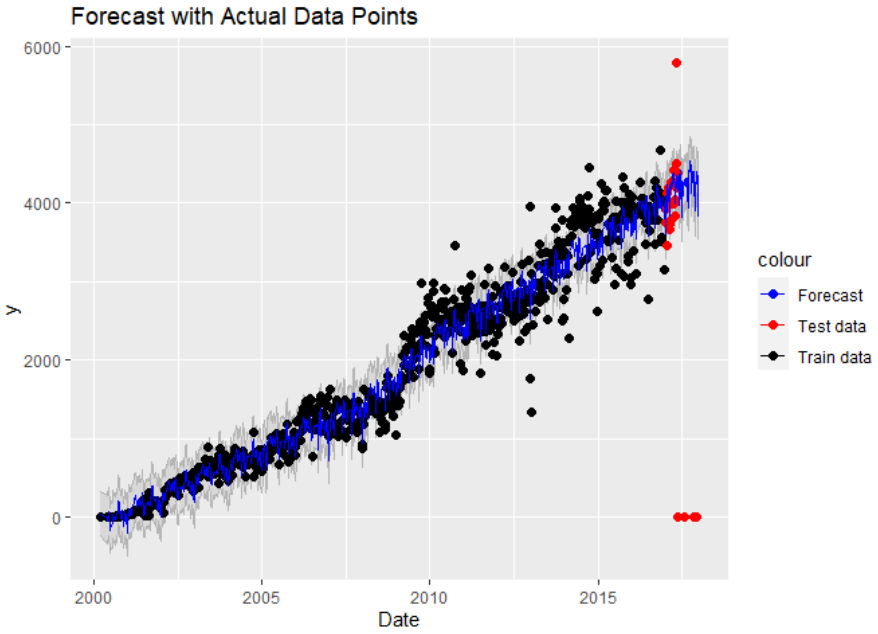

Figure 5: Prediction of 2017 Patent examiner production based on 2000 - 2016 historical Data and Accounting for Holidays. Actual production in red.

Measure	*Value*	*Measure*	*Value*
Mean Squared Error (MSE)	*2,872,742*	*Root Mean Square Error (RMSE)*	*1694.9*
Mean Absolute Error (MAE)	*962.8091*	*Mean Absolute Percentage Error (MAPE)*	*508.7*

3.1.3. Insights

Including holidays in our prediction model resulted in a small improvement in prediction wherein the error values all decreased by approximately 1.5%. However, this is not a significant improvement because problems with outliers and cutoff dates persist.

With this, our next course of action is to modify the date filtering for a cleaner dataset.

3.2. Setting the Cutoff Date to 2016

3.2.1. Background

The 2017 dataset had outliers which were proving to be problematic with making accurate predictions, so instead we plotted the 2016 data.

3.2.2. Plots

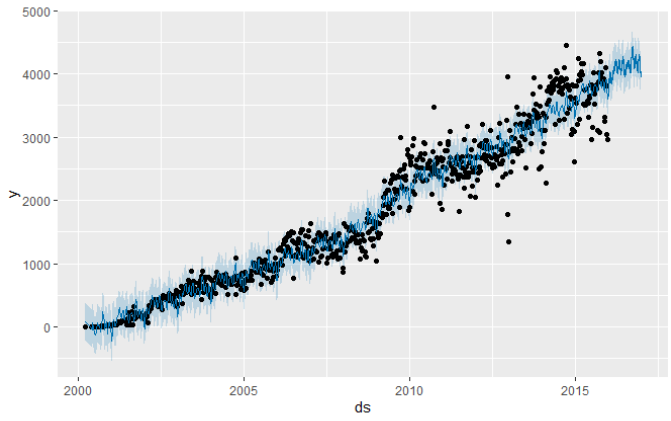

Figure 6: Prediction of 2016 Patent examiner production based on 2000 -2015 historical Data.

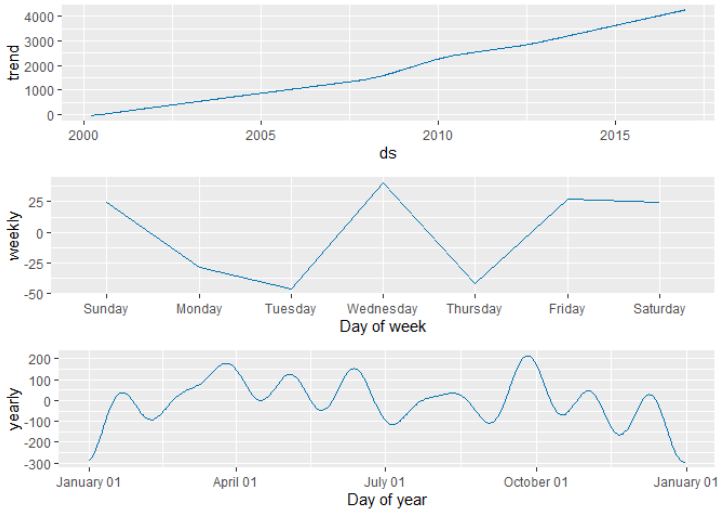

Figure 7: Component Plot of 2016 Holliday Predictions.
This Decomposes the data into Annual, Monthly, Day of Week and Holliday Trends.

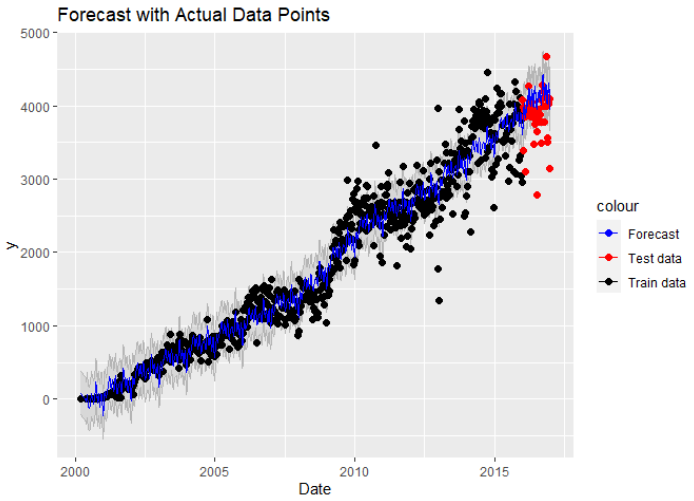

Figure 8: Prediction of 2016 Patent examiner production based on 2000 - 2015 historical Data. Actual production in red.

Measure	*Value*	*Measure*	*Value*
Mean Squared Error (MSE)	*185,234.2*	*Root Mean Square Error (RMSE)*	*430*
Mean Absolute Error (MAE)	*324*	*Mean Absolute Percentage Error (MAPE)*	*0.09*

3.2.3. Insights

Changing to 2016 data had a significant improvement both visually and in terms of the prediction values. Therefore, we decided to continue using 2016 data.

3.3. Changing to 2016 with Holidays

3.3.1. Background

Based on the 2016 data performance, we added holidays to the 2016 data model seeking to further improve model performance to get even higher accuracy predictions.

3.3.2. Plots

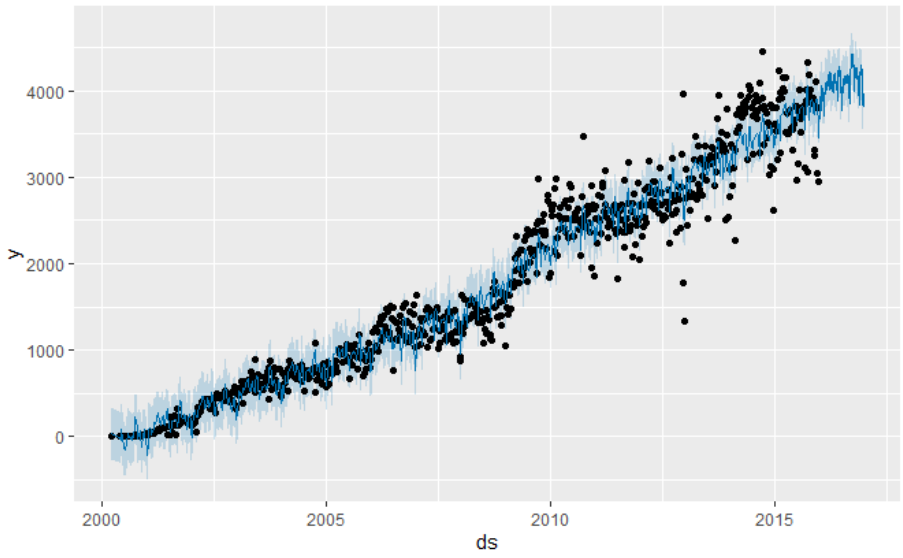

Figure 9: Prediction of 2016 Patent examiner production based on 2000 - 2015 historical Data.

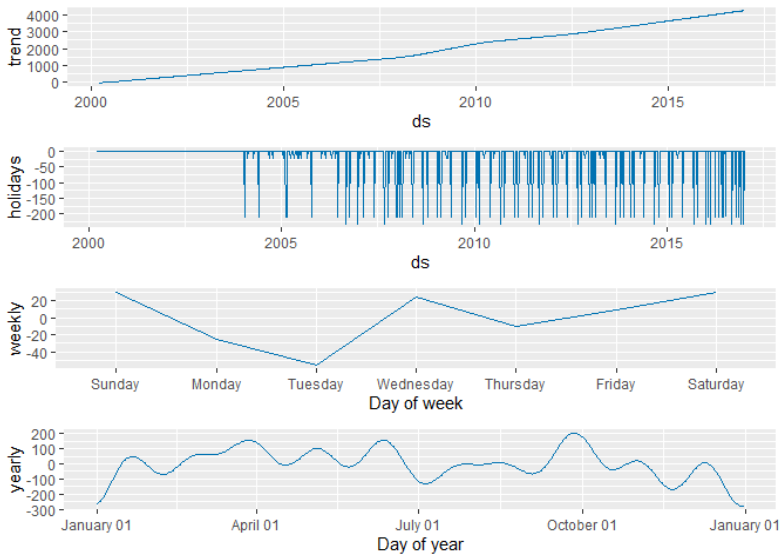

Figure 1010: Prophet Prediction of 2016 Weekly Production Numbers based on 2000 - 2016 Values Including Holidays.

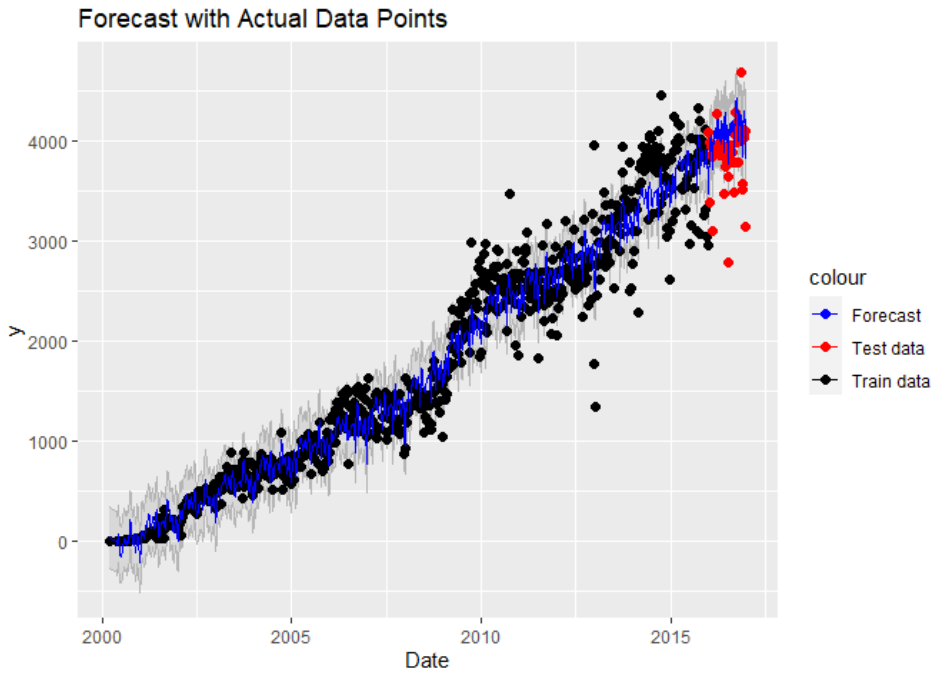

Figure 11: Prediction of 2016 Patent examiner production based on 2000 - 2015 historical data and Accounting for Holidays. Actual production in red.

Measure	*Value*	*Measure*	*Value*
Mean Squared Error (MSE)	*180,519.6*	*Root Mean Square Error (RMSE)*	*424.88*
Mean Absolute Error (MAE)	*324.6*	*Mean Absolute Percentage Error (MAPE)*	*0.09*

3.3.3. Insights

Visually the graphs appear to be quite similar and do not offer much improvement. Mathematically, however, the results show an improvement of 2.5%. We expect marginal improvement based on additional data and performance metrics.

We suspect adding Gender or data granularity will be needed for increased accuracy.

3.4. Gender Graphs

3.4.1. Background

As such, we attempted to evaluate the data based on gender to further break down the production rate based on examiner features. For this, we used the 2016 dataset with holidays and compared Male vs. Female data. Note: Examiners with no identified gender were excluded from the data.

3.4.2. Plots

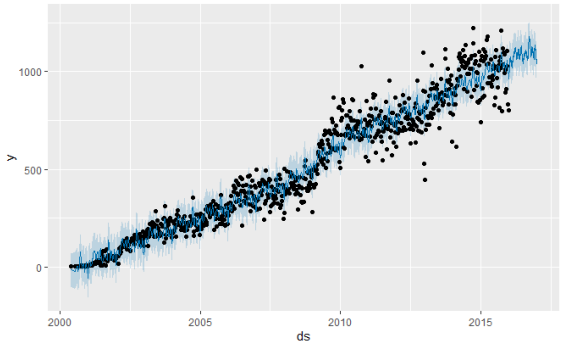

Figure 12: Prediction of 2016 Female Patent examiner production based on 2000 - 2015 historical Data.

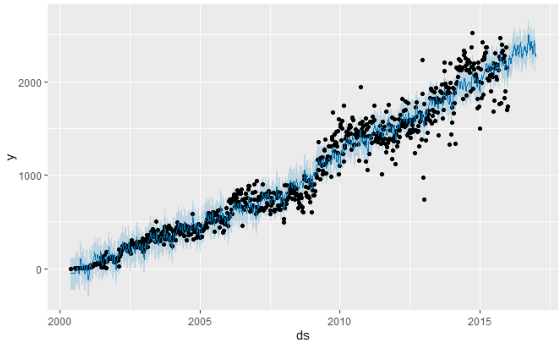

Figure 13: Prediction of 2016 Male Patent examiner production based on 2000 - 2015 historical Data.

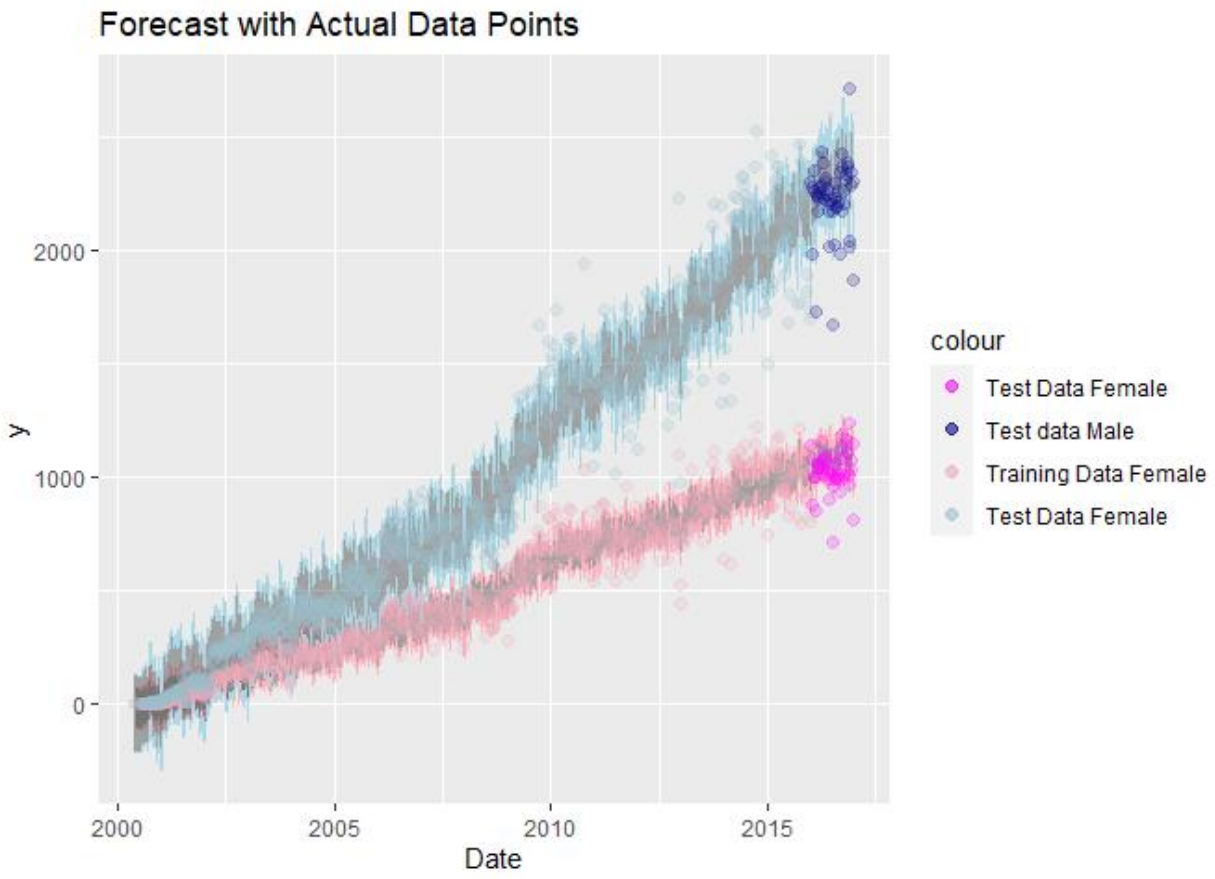

Figure 14: Prediction of 2016 Patent examiner production based on 2000 -2015 historical Data, separated out by gender and Accounting for Holidays. Actual production in red.

Table of Stats for Females

Measure	*Value*	*Measure*	*Value*

Mean Squared Error (MSE)	*48,468.04*	*Root Mean Square Error (RMSE)*	*220.15*
Mean Absolute Error (MAE)	*164.1*	*Mean Absolute Percentage Error (MAPE)*	*0.08*

Table of stats for Males

Measure	*Value*	*Measure*	*Value*
Mean Squared Error (MSE)	*13,213.5*	*Root Mean Square Error (RMSE)*	*114.95*
Mean Absolute Error (MAE)	*88.3*	*Mean Absolute Percentage Error (MAPE)*	*0.09*

3.4.3. Insights

Based on the gender plots, there is a clear difference between USPTO patent examiners who are men and women, in terms of productivity.

Separating these genders out increases accuracy significantly and offers more insights into the productivity of patent examiners as it relates to the number of applications they process. Therefore, separating the examiners by gender increases prediction accuracy for this model. This effect is notably apparent for female evaluators.

4.1. The Predictive Model

In our findings, we observed the 2016 Model with Holidays accounted for was the most accurate. However, it is important to note that predicting the number of patent applications is not an exact science, and there is always a risk of inaccurate predictions. Indeed, the predictive model is subject to the data quality and external factors that we do not have control over.

Moreover, in nature, we see logistic curves with natural processes such as in the population growth of animals, the spread of infectious diseases, and the growth of plants. Similarly, we expect the trend of increasing productivity as it relates to the USPTO's number of patent application processes will not be linear and perpetual. Rather, we suspect it will plateau over time like those natural logistic occurrences, because of a slowing global population growth and rate of patentable innovation.

4.2. A Business Case for Predicting Patent Applications

Revenue Generation: The USPTO charges fees for patent applications and predicting the number of applications would help estimate revenue for the year and enable the USPTO to plan for any potential shortfalls and adjust their fee structures as needed. (USPTO fee structure: https://www.uspto.gov/learning-and-resources/fees-and-payment/uspto-fee-schedule)

Resource Allocation for Higher Efficiency: Predicting the volume of applications can help the USPTO optimize resource allocation such as staff and budget to ensure that they are adequately prepared to handle the workload (i.e., avoid over or understaffing) which would drive cost savings and increased productivity.

Quality Assurance: Predicting the volume of applications could help the USPTO maintain high standards of quality by ensuring that examiners have enough time to thoroughly review applications.

Risk Management: Predicting the number of applications can help the USPTO identify risks such as budget or seasonal-like workload spikes and take appropriate measures to mitigate them.

Strategic Planning: Knowing the expected number of applications could help the USPTO plan for future initiatives such as hiring new examiners or expanding their services to new areas (i.e., growing TCs and adding art units or work groups).

Better Customer Service: Predicting the number of applications will help the USPTO manage customer expectations (i.e., take a proactive approach to processing delays) and provide better service to applicants.

4.3. Business Risks & Considerations

While there are several business value cases to predicting the number of patent applications, the USPTO must carefully consider the risks and trade-offs associated with predicting the number of patent applications, and have contingency plans in place to mitigate potential challenges. Predicting the number of patent applications is not an exact science, and there is always a risk of inaccurate predictions. This can lead to an over or under allocation of resources, which can impact efficiency, quality, and customer satisfaction.

Overestimating the number of patent applications can result in a budget shortfall, while underestimating can lead to inefficient use of resources and unnecessary costs. For example, considering rapid technological advancements, predicting the number of patent applications may be challenging in rapidly evolving fields such as artificial intelligence or blockchain technology, where the number of patent applications may surge unexpectedly.

Furthermore, unexpected events such as changes in laws or regulations, economic downturns, or geopolitical issues can affect the number of patent applications in any given year. The USPTO should be prepared to adjust their plans and resource allocation, accordingly, being flexible in their resource allocation in case the actual numbers deviate from their predictions.

Lastly, there may be pressure to process patent applications quickly to meet very high predicted volumes, potentially leading to a compromise in the quality of examination and granting patents that do not meet patentability standards.