Prediction of US Patent Approval Time

February 2023

Erik Pak
Data Scientist
Civil engineering (B.S.)

Masoumeh Dehghani
Data Scientist
Physicist (Ph.D.)

Robert Lewis
Data Scientist
Biomedical Engineering (B.S.)

- Strategic context
- Problem Definition
- Methodology, Analysis & Results
- Recommendations & Next Steps

Project | Strategic Context

What:

Predict the Approval time of patents in the US using a set of demographic variables

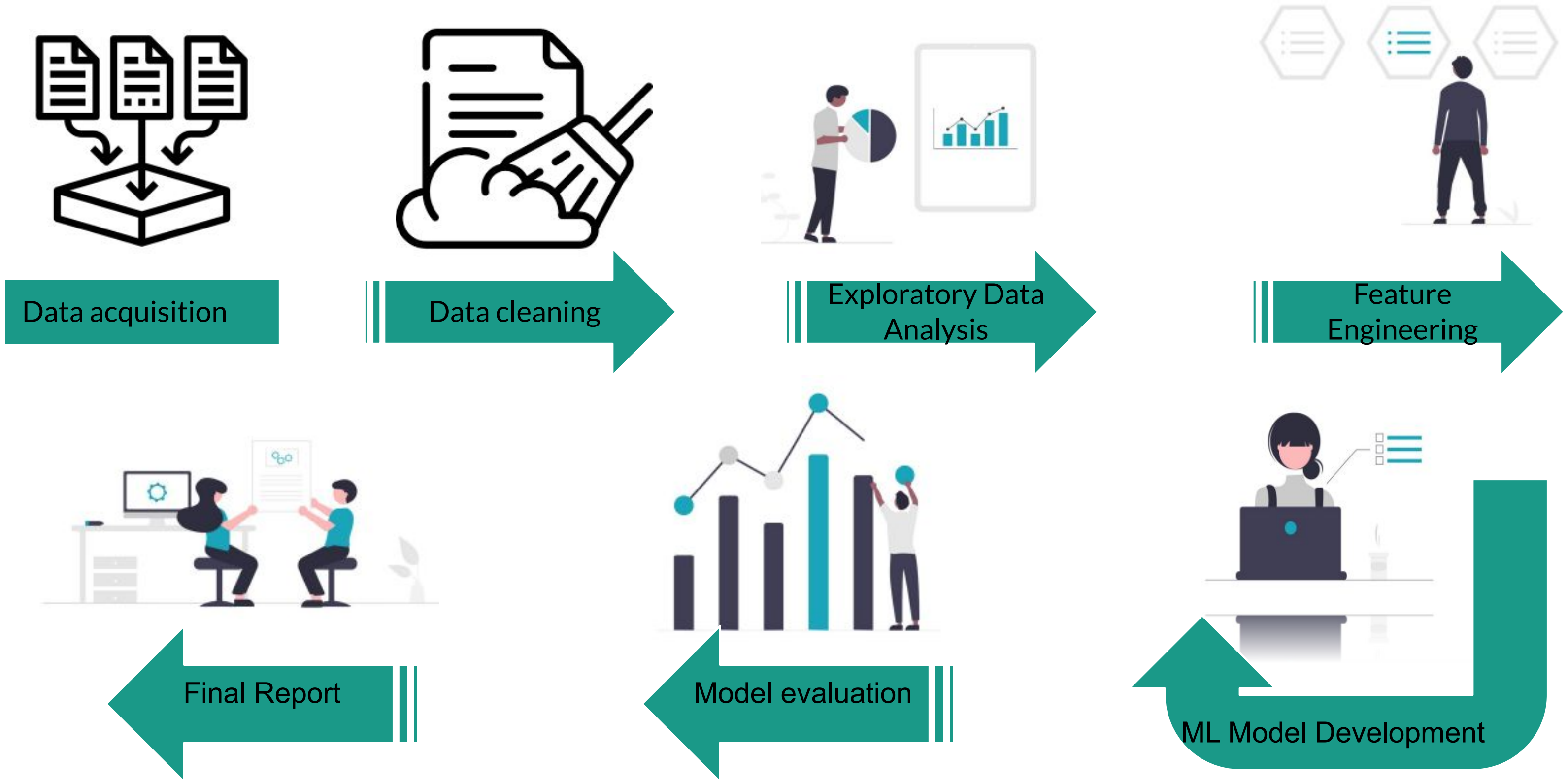
Why:

US patent approval remains a somewhat opaque process. If certain variables are determined to decrease approval time, it would be beneficial to inventors.

Project | Problem Definition

Currently there exists a lack of transparency in the patent approval process. However, expedited patent approvals would benefit both the inventor/patent holder and their customers. The need for improved patent approval time permeates every industry and academic research institution. The goal of this project is to try and determine what patent features will most often lead to an expedited patent approval process.

Project | Methodology



Project | Data acquisition

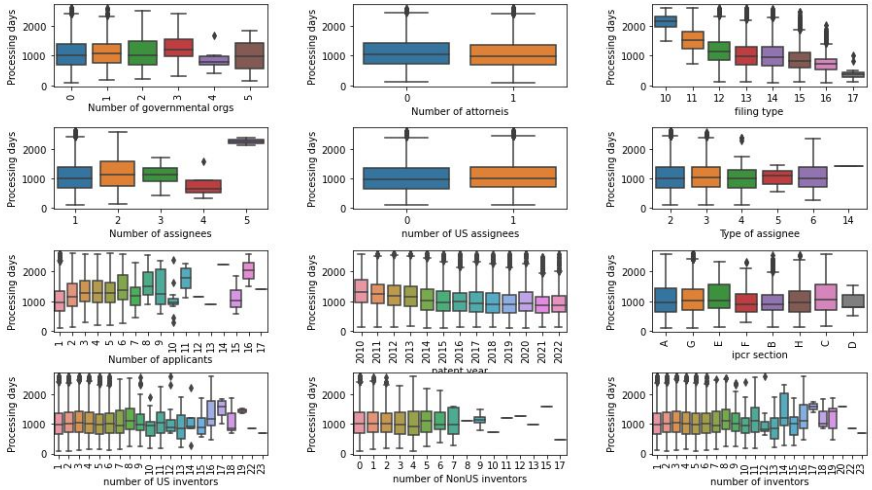
Data Source: PatentsView.org https://patentsview.org.

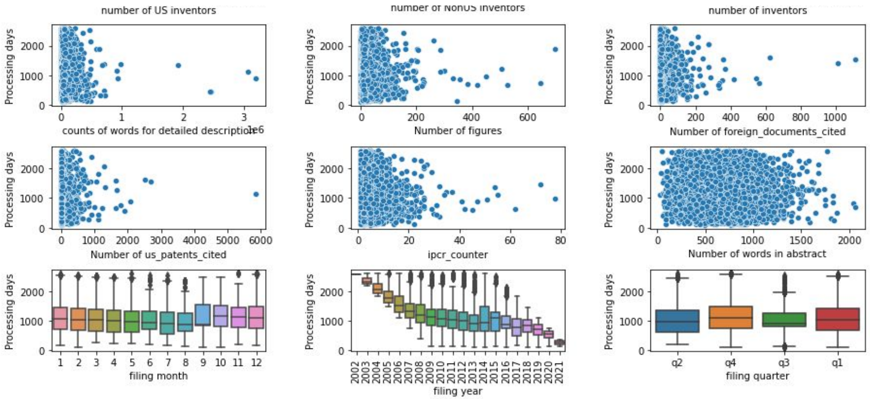
Features: 22 variable

Observation: 13,000 patent, cover year (2010-2022)

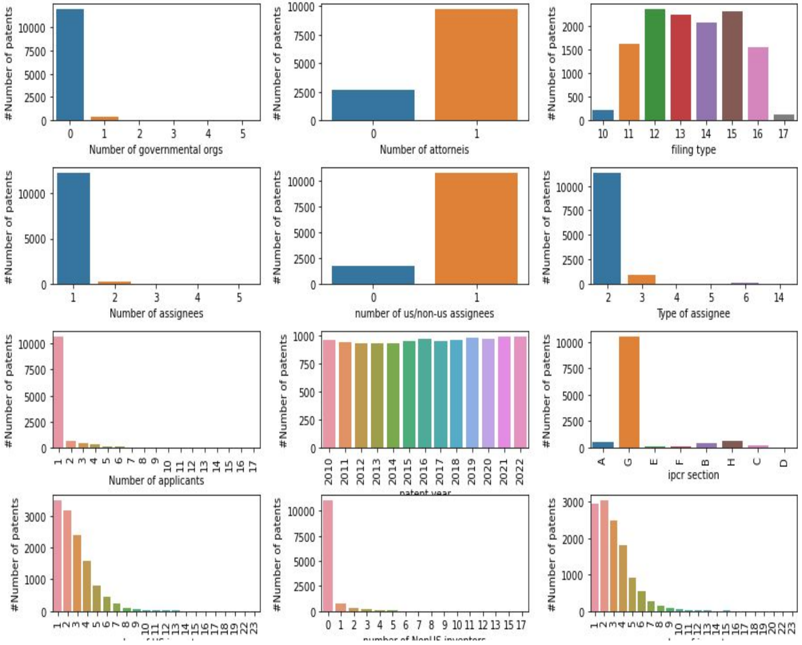
```
RangeIndex: 13000 entries, 0 to 12999
Data columns (total 22 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   patent_id                           13000 non-null  int64
 1   patent_date                         13000 non-null  datetime64[ns]
 2   count_gov_orgs                      13000 non-null  int64
 3   patent_detail_desc_length           11979 non-null  float64
 4   num_figures                         13000 non-null  int64
 5   attorney_count                      13000 non-null  int64
 6   filing_date                         13000 non-null  datetime64[ns]
 7   filing_type                         13000 non-null  object
 8   n_assignees                         12246 non-null  object
 9   assignee_isUS                       12170 non-null  object
 10  assignee_type                       12185 non-null  object
 11  n_applicants                        9010 non-null   object
 12  patent_num_foreign_documents_cited  13000 non-null  int64
 13  patent_num_us_patents_cited         13000 non-null  int64
 14  patent_year                         13000 non-null  int64
 15  ipcr_section                        13000 non-null  object
 16  ipcr_counter                        13000 non-null  object
 17  inventor_US                         13000 non-null  int64
 18  inventor_NonUS                      13000 non-null  int64
 19  inventor_counter                    13000 non-null  int64
 20  patent_abstract_counter             13000 non-null  object
 21  patent_processing_days              13000 non-null  int64
dtypes: datetime64[ns](2), float64(1), int64(11), object(8)
memory usage: 2.2+ MB
```

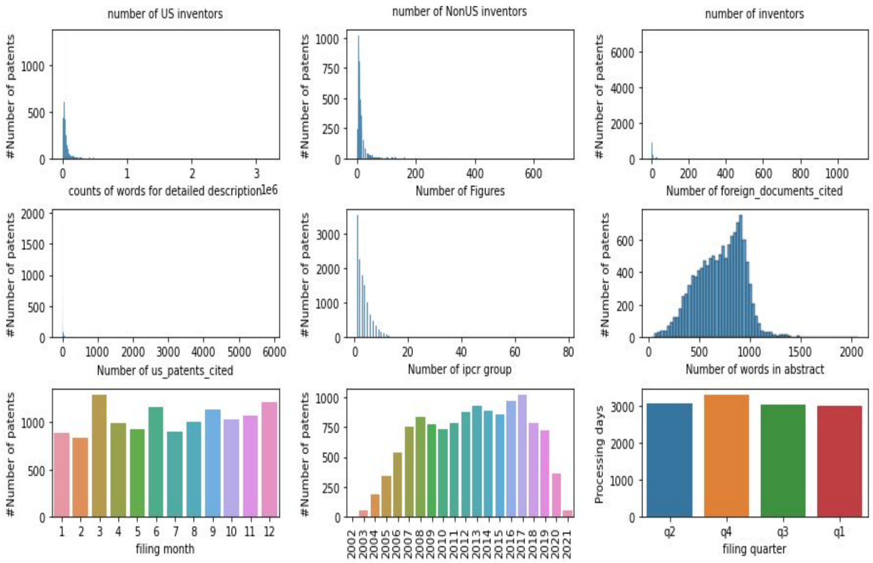
Project | Exploratory Data Analysis





Project | Exploratory Data analysis





Project | Data cleaning & Feature engineering

Cleaning Data

Number of patents: 12,439
Number of variable: 13

```
RangeIndex: 12439 entries, 0 to 12438
Data columns (total 13 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   patent_processing_days              12439 non-null  int64
 1   count_gov_orgs                      12439 non-null  int64
 2   patent_detail_desc_length           12439 non-null  int64
 3   num_figures                         12439 non-null  int64
 4   attorney_count                      12439 non-null  int64
 5   n_applicants                        12439 non-null  int64
 6   patent_num_foreign_documents_cited  12439 non-null  int64
 7   patent_num_us_patents_cited         12439 non-null  int64
 8   ipcr_counter                        12439 non-null  int64
 9   inventor_counter                    12439 non-null  int64
 10  ipcr_section                        12439 non-null  object
 11  filing_type                         12439 non-null  int64
 12  filing_quarter                      12439 non-null  object
dtypes: int64(11), object(2)
```

Scaling numerical features

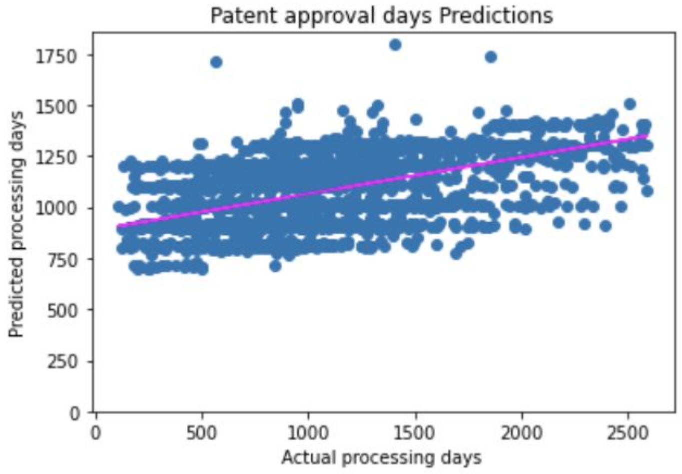
Encoding categorical features

Project | Math Models

1. Linear Regression
2. Random Forest
3. XGBoost

Project | Model Evaluation

Linear Regression:

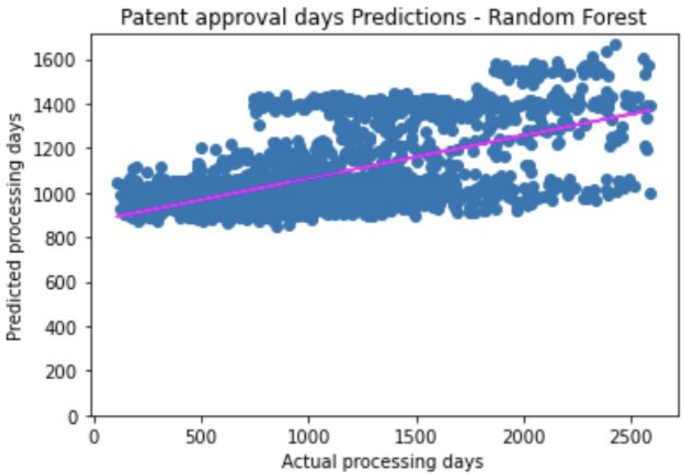


LR is a common technique used to predict a continuous target variable by finding a linear relationship between it and the feature set of independent variables

MSE: 204370.87074407292
RMSE: 452.0739660100689
R2: 0.23885213786853787

Project | Model Evaluation

Random Forest:

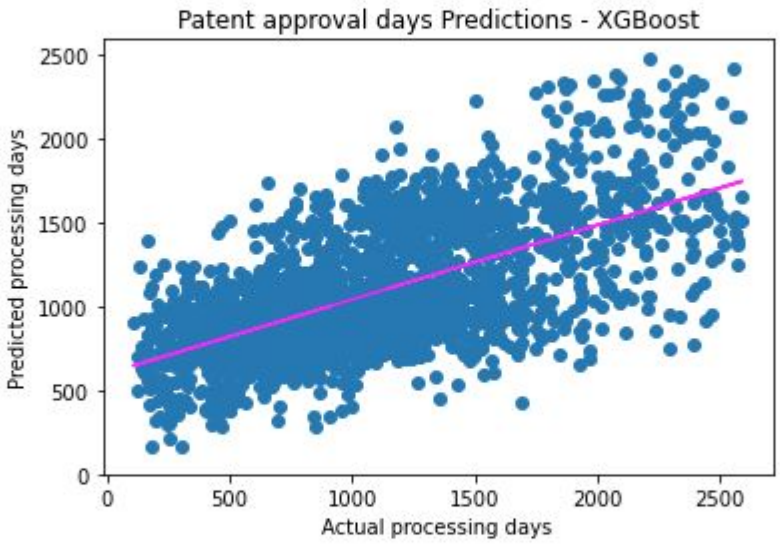


In Random Forest Regression, features are chosen at random and used to create many decision trees. These 'trees' are then averaged together to create a more accurate prediction.

MSE: 194586.84180557213
RMSE: 441.1199857244876
R2: 0.27529124820974693

Project | Model Evaluation

XGBoost*:



In XGBoost, weights are given to each independent variable and a decision tree is made. The weights of the wrong predictions are increased and fed into a sequential decision tree. The trees are then brought together to create a more accurate prediction.

```
MSE: 160289.75462993115
RMSE: 400.3620294557554
R2: 0.4030254722018057З
```

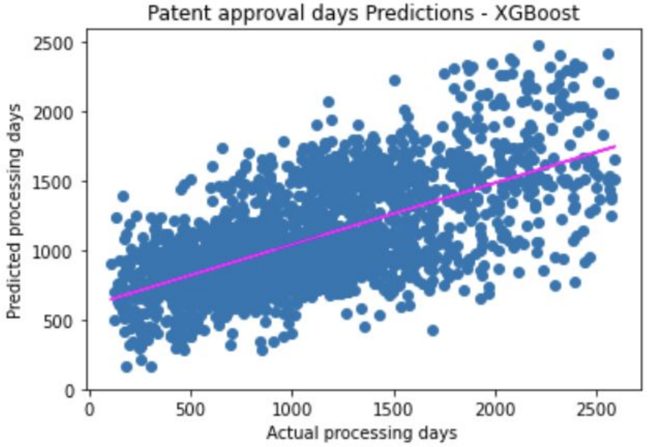
Project | Final Model

XGBoost:

It provide the best predictive power

Optimization:

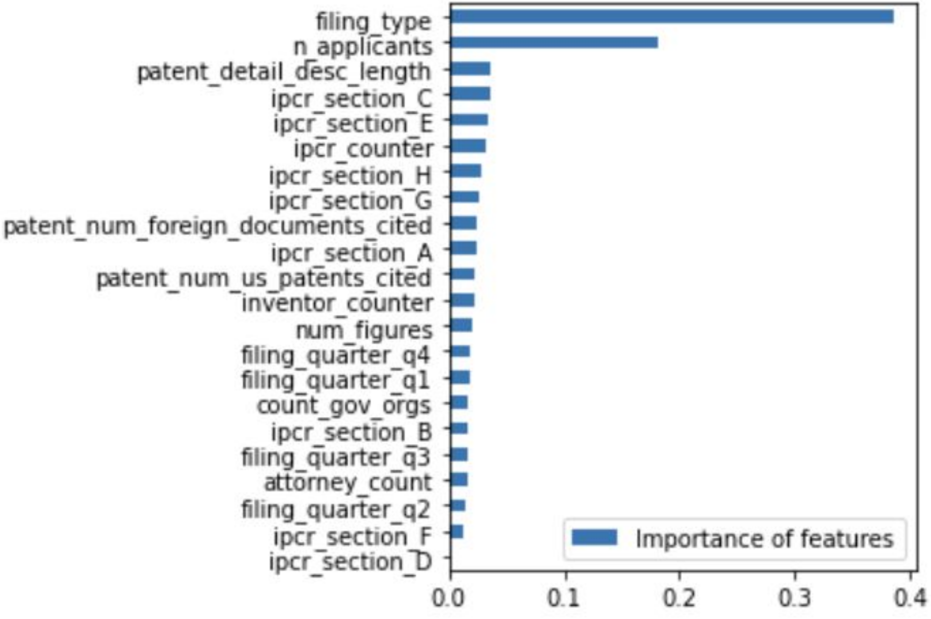
GridSearchCV




```
MSE: 150348.56689095014
RMSE: 387.74807142131624
R2: 0.4400498963138332
```

Project | Analysis & Results

Feature importance



Features of highest importance, from the demographics variables we chose were:

1. Filing Type
2. Number of Applicants

Project | Recommendation

Data & Technical Next steps:	- Test model on more data - Continuous development of ML model - Add / Engineer new variables
Business Next Steps:	- Model integration and business requirements - Deliver AI training and education to users