Fighting with the Nature of Patent Data: Predicting Patent Citations

Enver Ferit Akın

CSSM 502 Final Project

Instructor: Mehmet Fuat Kına

January 2023

KOÇ
ÜNİVERSİTESİ
GRADUATE SCHOOL OF SOCIAL
SCIENCES AND HUMANITIES

1 Introduction

"The main object of all science is the freedom and happiness of man."

Thomas Jefferson, the first Commissioner of the U.S. Patent Office

Innovation plays a crucial role in a society's economy and social development. It drives productivity, creates new industries, and usually improves the quality of life. Understanding the innovation process and its dynamics is essential for policymakers and researchers to develop effective policies and strategies for fostering innovation. In my master's thesis, I am working on a topic that integrates innovation and migration. It is an empirical project where I estimate the causal effect of international graduate students on universities' innovative output. To measure innovative output, I use not only the number of patents of a university but also the quality of patents. Patent citations are essential for measuring a patent's impact and quality, so I use patent citation data. Also, many papers, especially in the economics of innovation and industrial organization, use patent citations as a metric for patent quality. However, the nature of patent data complicates these analyzes. The central challenge is the truncation of the patent citation data. Newly filed or granted patents have very few citations, so using these patents when measuring patent quality may cause biased inferences (citation bias.) Some methods offer solutions to this problem. I will refer to these methods in the literature section. The inadequacy of these methods is discussed in some papers. Thus, this project aims to develop a machine-learning model that can accurately predict the number of total citations a patent would receive while considering the truncation of the patent citation data.

The structure of this paper will be as follows: I discuss some of the articles that focus on the truncation problem of patent citation data in the literature section. The data section describes the datasets I used in my models. I also explain the data processing procedure. In the methodology sec-

tion, I present machine learning algorithms' details and results. The conclusion section summarizes the findings. Besides, the conclusion section discusses the limitations of this research and potential research questions.

2 Literature

Hall, Jaffe, and Trajtenberg (2001) propose an adjustment method for the citation bias in their seminal paper. Their methodology is called the "quasi-structural" adjustment. They estimate the following functional form,

$$f_k(L) = \alpha_0 + \alpha_t + \alpha_k - log(C_{kst}/P_{ks}) \tag{1}$$

Where $\alpha_j = log(\alpha'_j)$, and $f_k(L)$ indicates the citation-lag distribution, C_{kst} is the total number of citations to patents in year s and technology class k, coming from patents in year t, P_{ks} is the total patents in field k in year s. α parameters are some metrics for citation intensity for a given field or year. Then, estimated $f_k(L)$ can be used to adjust the truncation problem. A more detailed discussion can be found in their paper. Although this method is widely used, Lerner and Seru (2017) claim that this adjustment method may not solve citation bias. Instead, they argue that machine learning could solve the citation bias. They use machine learning algorithms to predict the total yearly citations that a firm's patents will receive. They benefit from firm-level data to create the features of their models. The main difference between our method and their method is that they predict the number of citations at the firm level while we predict it at the patent level. Also, they predict yearly citations while we predict the total citations a patent would receive.

3 Data

I used the United States Patent and Trademark Office's (USPTO) disambiguated assignee, patent citation, patent information, patent classification, and patent application datasets in this project.[1] Only patents filed or issued in the USA are included in these datasets. The common element of these datasets is the unique patent ids. The disambiguated assignee dataset contains the assignee information (the assignee is the patent owner, such as a firm, a university, or a public institution) of each patent. Since there are different variations of the name of an assignee, the USPTO uses a disambiguation algorithm to overcome this problem. Patent citation data contains citing and cited patents. Each row in this dataset represents a citation. There are approximately 129 million rows in this dataset. Note that these citations are citations made to US-granted patents by US patents. For instance, this dataset does not include a citation to a US patent from a patent granted in China. Patent classification data includes International Patent Classification (IPC) class codes. Patent application data contains the application year information of each patent.

Given the large size of the datasets, I performed random sampling by selecting 10,000 random assignees and their patents. I then used these patents to select their citations, resulting in a sample size of 1,524,886 citations and 86,492 unique cited patents. I performed several data cleaning and preprocessing steps to prepare the datasets for analysis. This process included removing duplicate entries and merging the various datasets to create a comprehensive dataset. I also created new variables that are relevant to our research. I summarized these processes in Figure 1. Also, the data processing codes are in the "Final_Data_clean.ipynb" file on my GitHub page.[2] The final dataset used in this project includes information such as the patent id, assignee name, assignee type, patent classification, application and grant year, number of claims, class size, backward citations, novelty score (calculated in the "Final_Project_Models.ipynb" file), circulation, pendency, citation counts,

[1] The datasets can be found at https://patentsview.org/download/data-download-tables.
[2] https://github.com/efakin/Predicting-Patent-Citations

and classification (which shows whether the total number of citations of a patent is high, medium, or low. See "Final_Project_Models.ipynb".) The input dataset (model_input_final) contains 75,175 unique patents. I used to train and evaluate machine learning algorithms. It is worth noting that the data used in this project represents the patents in the USPTO dataset, and one should interpret any conclusion and prediction made in this project accordingly.

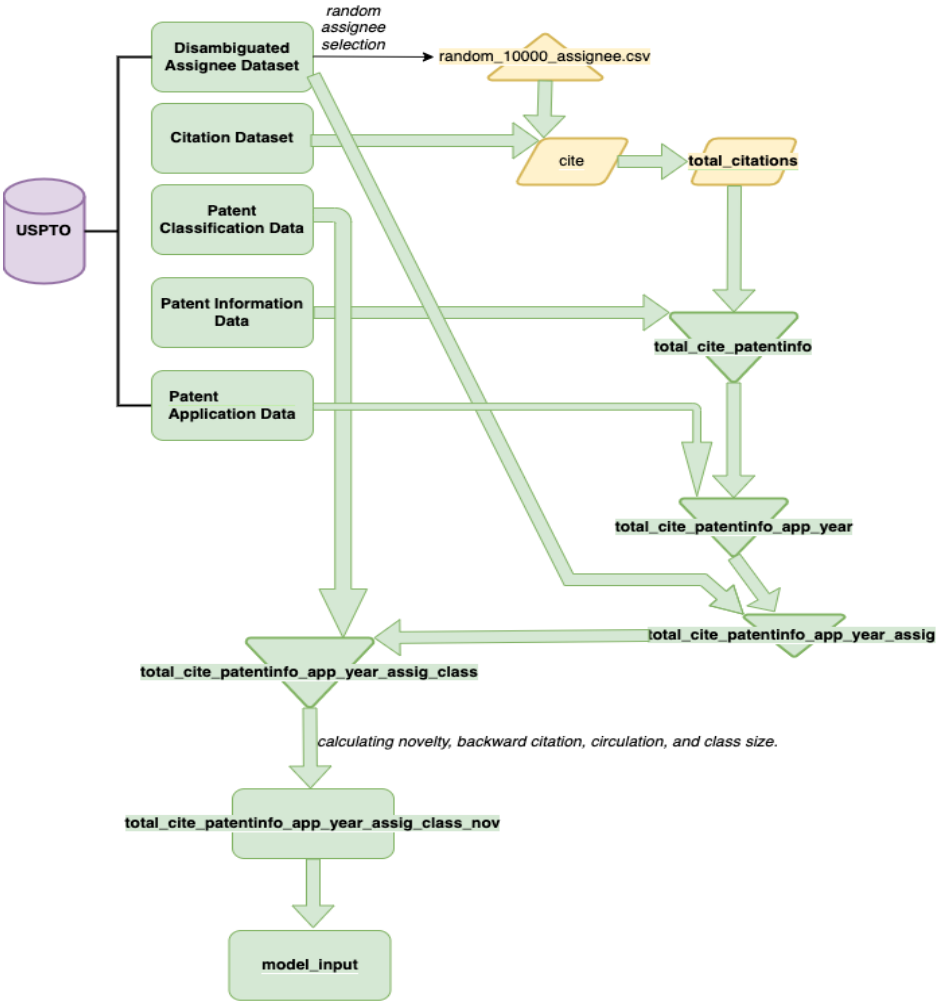


Figure 1: Data Processing Flowchart.

4 Methodology & Results

To predict the total number of citations a patent will receive, I modified the model input data in the "Final_Project_Models.ipynb" file. Our patent citation dataset covers the years between 1975 and 2022. The prediction problem was approached as both a classification and a regression problem. Before discussing the models, let us observe the distribution of patent citations by application year. Figure 2 strikingly illustrates the truncation of the patent citation data. Truncation for a patent's citations always exists. However, truncation becomes more severe for newly applied or granted patents. In our data, especially after 2002, there is a decline in the number of citations per patent. Therefore, the training data used for the model is the total number of citations of 44,707 patents applied between 1975 and 2002, i.e., the training data contains total number of citations these patents received until 2022 (total lifetime citations.)

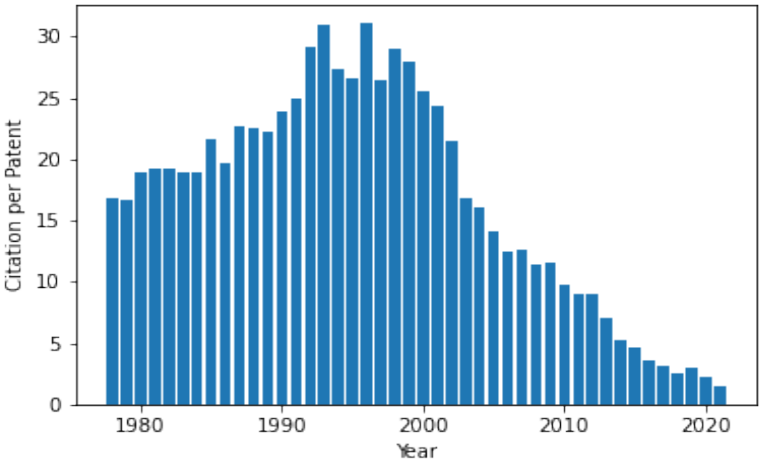
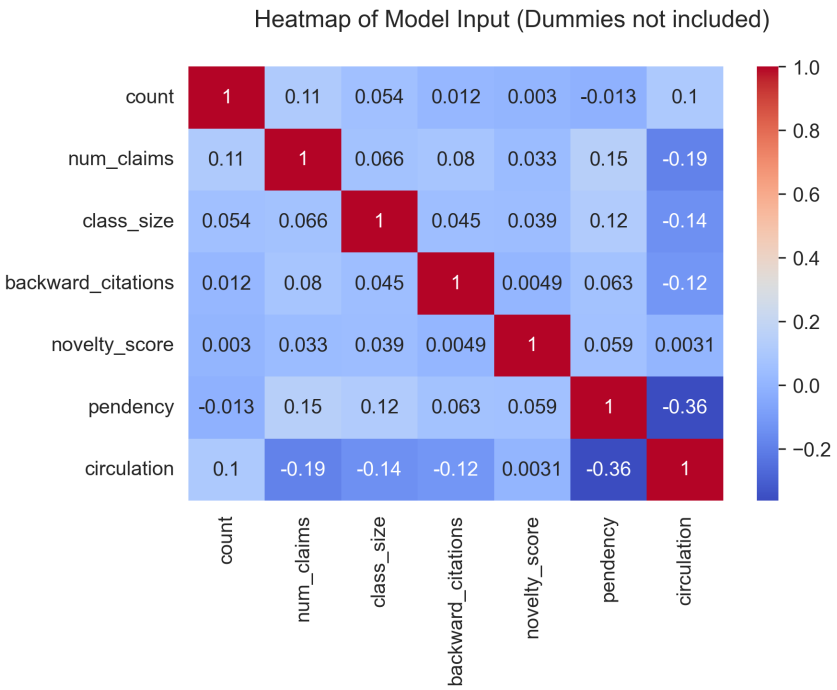


Figure 2: Citation per Patent by Application Year.

At the beginning of the project, I planned to predict the number of citations a patent would receive in a specific year. To that end, I created panel data with the patent, year, and yearly citations of the patent. However, the panel data was unbalanced since the application years of these patents

are different. For example, a patent filed in 1998 does not have any information before 1998. Since unbalanced panel data is problematic in machine learning, I decided to collapse the data at the patent level and use the total number of citations a patent received until 2022 as the target variable. The number of patent claims, application year, assignee type, class size, backward citations, novelty score, circulation, pendency, dummies for grant year, and dummies for patent classification are used as features. The features that I created are backward citations, class size, novelty score, circulation, and pendency. Backward citations are the number of citations that the cited patent made. Class size is the total number of patents in a patent class. The novelty score is the number of citations from outside the patent's class divided by the total citations of the patent. Circulation is the difference between 2022 and the grant year of the patent. Pendency is the difference between the grant year and the application year of the patent. The following heatmap shows the correlations between variables.



4.1 Classification Problem

I approach the prediction problem as multiclass classification and binary classification. Patents were classified as less cited, medium cited, and highly cited. The classification was done based on the percentiles of the total number of citations, with 0 meaning a patent is less cited (below the 50^{th} percentile), 1 meaning medium cited (between 50^{th} and 75^{th} percentiles), and 2 meaning highly cited (above the 75^{th} percentile of the total number of citations). In the binary classification setup, I classified patents as less cited and highly cited based on whether their total citations are below or above the 75^{th} percentile of total citations. Several popular machine learning models were used to classify the patents, including KNN, SVM, Decision Trees, Random Forest, Artificial Neural Networks, and XGBoost. In the binary classification setup, most of the models' performance increased. However, in either case, XGBoost was the best model by far. F1 scores and accuracy scores are used for the evaluation of the models. Table 1 and Table 2 summarize the models' accuracy and weighted average F1 scores. Classification reports can be found in the Jupyter notebook on my GitHub page.

Table 1: Accuracy and F1 Scores of Models in Multiclass Classification Setup.

Model	**Accuracy**	**Weighted Average F1 Score**
KNN	0.46	0.42
SVM	0.50	0.42
Decision Trees	0.57	0.64
Random Forest	0.52	0.39
Artificial Neural Networks	0.51	0.61
XGBoost	0.79	0.80

Table 2: Accuracy and F1 Scores of Models in Binary Classification Setup.

Model	**Accuracy**	**Weighted Average F1 Score**
KNN	0.72	0.75
SVM	0.75	0.80
Decision Trees	0.78	0.73
Random Forest	0.76	0.66
Artificial Neural Networks	0.74	0.77
XGBoost	0.87	0.86

In the multiclass classification setup, the worst performing model is KNN with accuracy of 0.46. This is not a good classifier, but it is still better than random classifying, i.e., with accuracy of 0.33. I used 10-fold CV for model tuning for some models. However, I could not use a 10-fold CV for the other models since it took a lot of time. Instead, I used a 3-fold CV. Especially in the multiclass classification setup, models other than XGBoost performed poorly. For some of the poorly performed models, I checked for overfitting by comparing training and test accuracies. There was no big difference between them, so overfitting is not a problem in the multiclass and binary classification cases. On the other hand, models other than XGBoost have minimal F1 scores for class 1, which is medium cited. For instance, Random Forest's precision and recall scores for class 1 are 0, which means it is a weak classifier.

4.2 Regression Problem

I also approached this problem as a regression problem. I tried to predict the exact number of total citations that a patent would receive. I used Lasso, Ridge, Random Forest, KNN Regressor, Elastic Net, Decision Trees, and XGBoost. RMSE is a widely accepted evaluation metric in regression problems, so I used RMSE as the evaluation metric. Table 3 summarizes the performances of each model. Again, XGBoost stands out in this case.

Table 3: Model Performances in Regression Setup.

Model	**RMSE**
Lasso	47.33
Ridge	47.33
Random Forest	45.35
KNN	48.52
Elastic Net	47.89
Decision Trees	47.94
XGBoost	43.33

5 Conclusion & Future Work

In this project, I aimed to create a machine-learning model to predict the total citations a patent will receive by using the number of patent claims, application year, assignee type, class size, backward citations, novelty score, circulation, pendency, dummies for grant year and patent classification as features. In both classification and regression setups, XGBoost is the best-performing model. My main goal is to use these predictions when designing a metric that measures patent quality in my future research. I will benefit from the predictions made by XGBoost. In this way, I hope to minimize the problem of citation bias arising from the nature of patent data.

Some factors such as the complex and dynamic nature of technological innovation, the large number of patents, and the patent system's complexity make predicting patent citations challenging. Also, one drawback of my analysis is that I only used the patents with at least one citation, but using patents with 0 citation would increase the quality of the analysis. Despite all this, I believe there are fruitful research questions in this area. First, textual features of the patent (such as title, abstract, and claims) can be extracted from the patent using natural language processing. This information has the potential to explain the patent quality. Second, a network analysis of patent citations can be done. Then, network-based features like co-authorship and the degree of knowledge spillovers between assignees may explain the patent citations. Finally, using the inventor's information can be beneficial in explaining patent citations.

References

[1] Hall, B., Jaffe, A., & Trajtenberg, M. (2001). The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. *NBER Working Paper.*

[2] Lerner, J., & Seru, A. (2017). The Use and Misuse of Patent Data: Issues for Finance and Beyond. *The Review of Financial Studies.*