Predicting Forward Citations for Patents

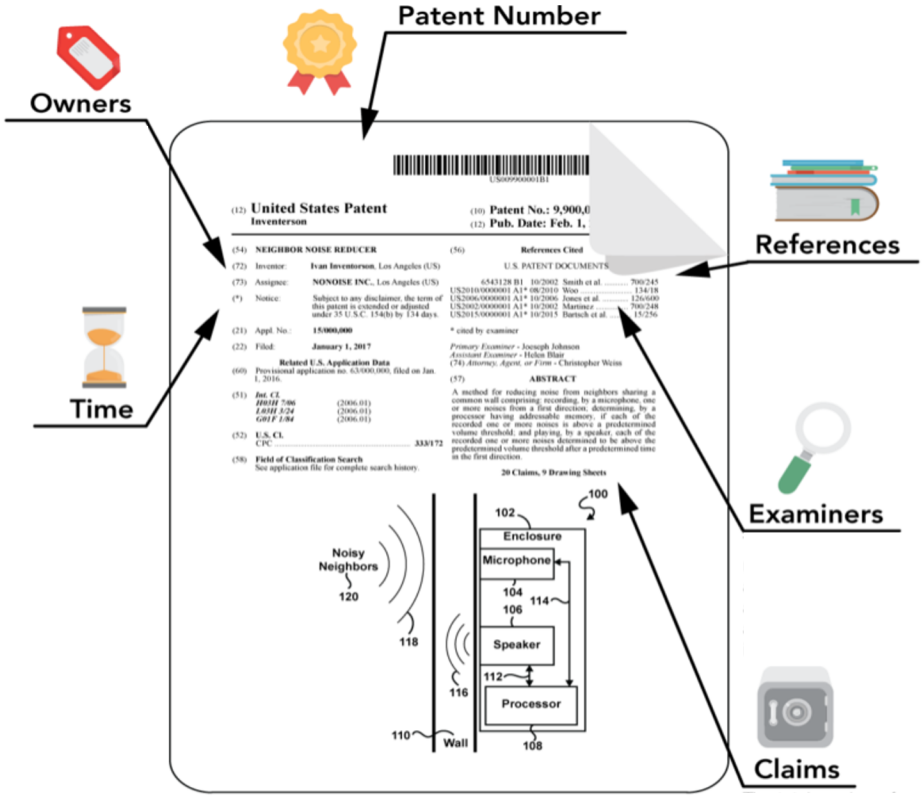
Akshay Pakhle (avp2131)
Harguna Sood (hs3159)
Siddhant Shandilya (ss5919)
Siddhanth Vinay (sv2609)
Swarna Bharathi Mantena (sm4776)

Industry Mentor
Eric Kang
Vice President - Quantitative Investment Strategies at Goldman Sachs

DSI Mentor
Kriste Krstovski
Associate Research Scientist and Adjunct Assistant Professor

Problem Statement

- **Problem definition:**
  - **Predict the number of citations a patent may receive**
- **Assumption:**
  - **A patent's value can be quantified by its forward citations**
- **Motivation:**
  - **Aid investment decisions of Goldman Sachs by evaluating organizations' intellectual properties**
  - **Value of patents filed by an organization can be used as a proxy for evaluation**
  - **Quantifying a patent's value is a hard problem**



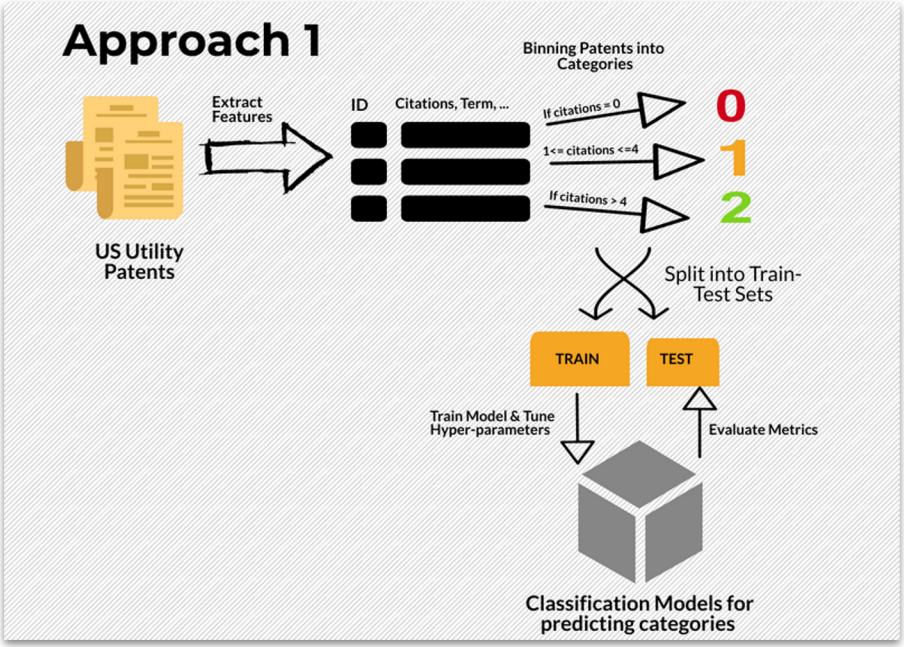
Problem Statement

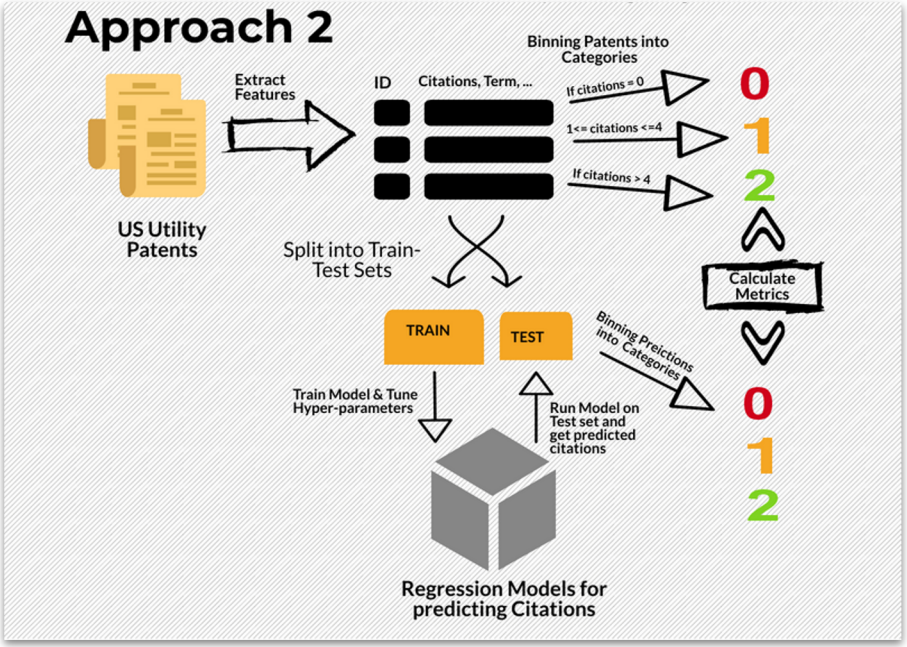
- **Solution:**
  - **Build predictive models to estimate forward citations for a patent**
  - **Investigate and interpret the impact of features on model prediction**
- **Past research has mostly focused on:**
  - **Forward citation prediction for research papers**
  - **Whether a patent will receive a forward citation**

Overview

Approach 1: Classification of patents into categories directly, based on number of citations (classification)

Approach 2: Predicting the number of citations for patents and then binning them (regression + binning)





Patent Terminology

- Claim - A claim defines the subject matter that is being protected by a patent
  - Independent claims - Standalone claims that contain all the information necessary to define an invention
  - Dependent claims - Claims that are dependent on other claims in the same patent
  - Exemplary claims - Claims that serve as an example to illustrate the meaning behind the patent
- Inventor(s) - The individual(s) who contributed to the claims of a patent
- Forward citations - Citations a patent receives
- Backward citations - Citations made by a patent
- Grant lag - Time between when a patent was filed and when it was granted
- Utility patent - Type of patent that covers the creation of a new or improved product, process, or machine

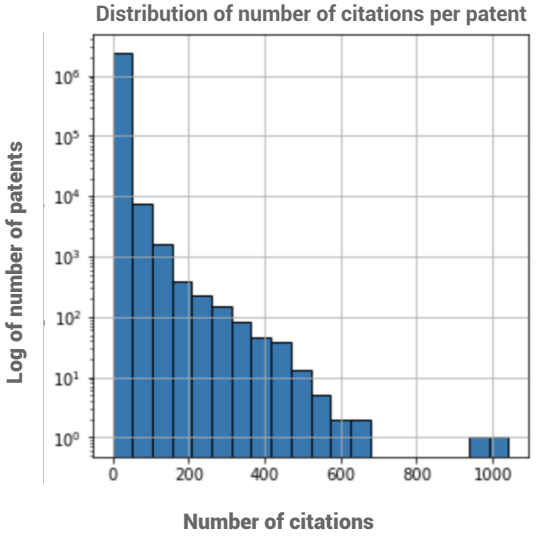
Data

- **Source - https://www.patentsview.org/web/**
- **We consider the following patents:**
  - **Utility type**
  - **Filed by organizations in the US**
  - **Granted after 2002**
- **For each patent, number of citations in the initial 5 years is calculated (to remove time bias)**
- **Resultant dataset contains 2.35 million patents**

Data

- **Distribution of citations is highly skewed; ~80% of patents have less than 5 citations**
- **Extremely hard to get a point estimate for this distribution**
- **Problem is converted to the following classification problem:**

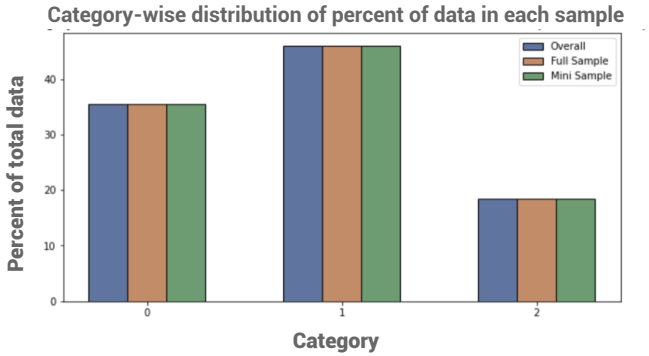
Category	Number of citations	Percent of data
0	0	41%
1	1-4	43%
2	5+	16%

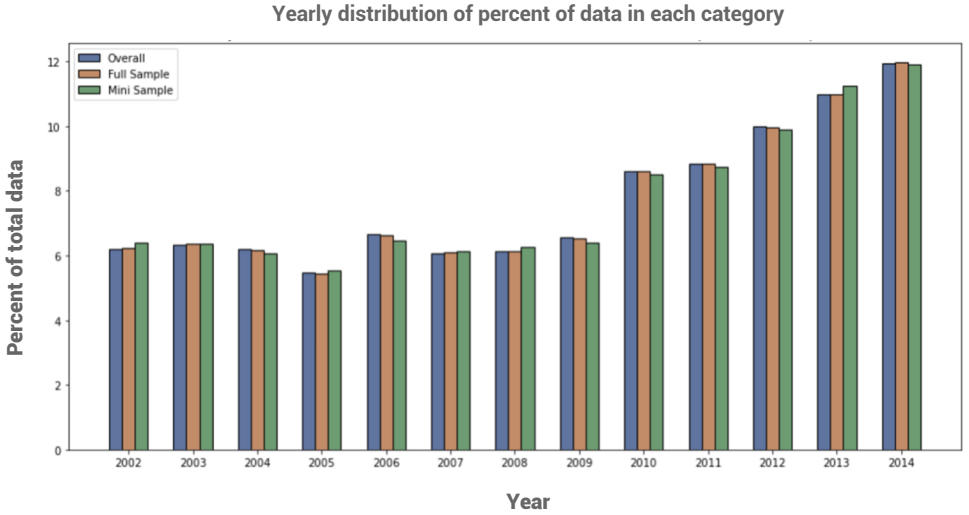


Quantile	Number of Citations
25	0
50	1
75	3
90	8
95	12

Sampling

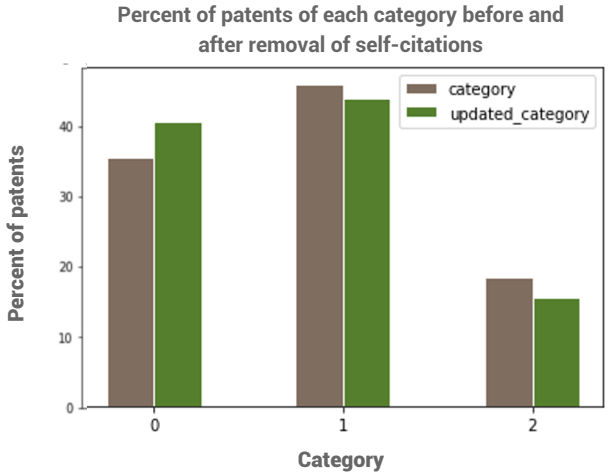
- **Performed stratified sampling to create two sample datasets:**
  - **Full (306K patents - for reporting results)**
  - **Mini (30.6K patents - for tuning purposes)**
- **Percent of each patent category is maintained in each sample**
- **Percent of patents distributed over all the years is maintained in each sample**





Bias Removal

- **Self-citations - citations by the same organization filing a patent**
- **Self citations are removed as they:**
  - **may induce bias**
  - **add less value to the organization**



Raw Features

- **Features directly available to us in our data:**
  - **Number of sections and subgroups under CPC* and IPCR****
  - **Number of total, independent and exemplary claims**
  - **Number of backward citations**
  - **Number of inventors for a patent**
  - **Grant lag**
  - **Number of sheets and figures**
  - **A flag indicating government interest**

* CPC = Cooperative Patent Classification;
** IPCR = International Patent Classification Reform

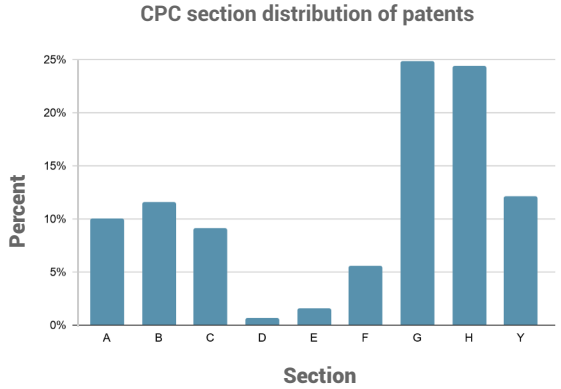
Derived Features

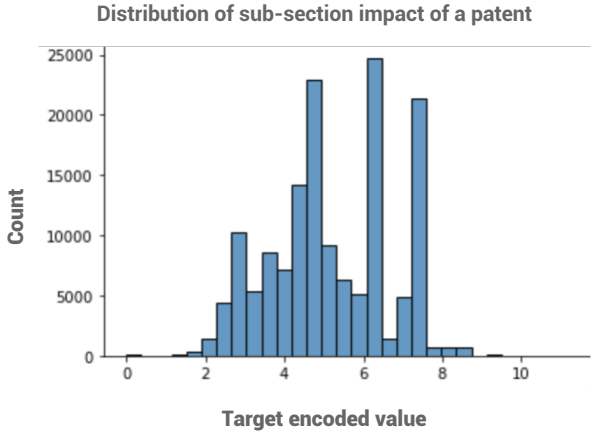
CPC sections

- One-Hot-Encoded values indicating the section of a patent
  - A = Human Necessities,
  - B = Performing Operations; Transporting,
  - C = Chemistry,
  - D = Textiles,
  - E = Fixed Constructions,
  - F = Mechanical Engineering,
  - G = Physics,
  - H = Electricity,
  - Y = Other

CPC sub-section impact

- Replaces each subsection by the average citation count of the subsection a patent belongs to
- Calculated average of all subsections a patent belongs to

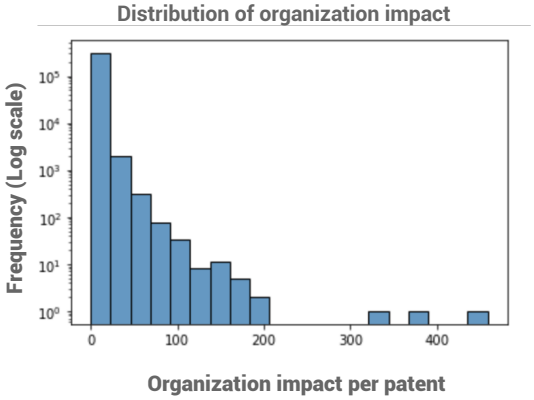


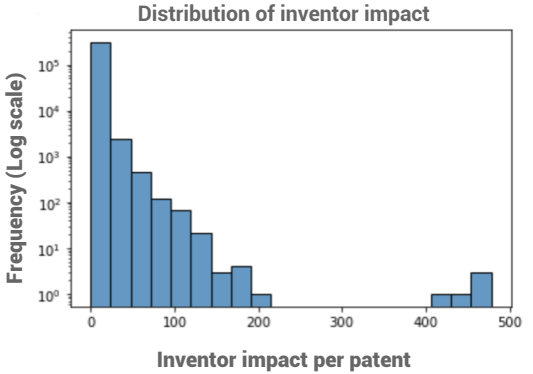


Derived Features

Organization/Inventor Impact

- **Captures the average number of citations per patent that the organizations/inventors filing a patent received**
- **Both features capture prior information regarding the quality of the organization/inventor**





Text Based Features

Rare and Frequent Words

- % of words in the abstract having count:
  - less than 85th percentile (rare)
  - between 90th and 95th percentiles (frequent)

Percentile	Value
50	14
85	310
90	339
95	1101
100	242548

Rare Words

Frequent Words

Text Based Features

Topic Modeling

- Trained an LDA model on independent claims of patents
- Document-topic distribution of patents used as features
- Topic-word distributions useful for interpretability

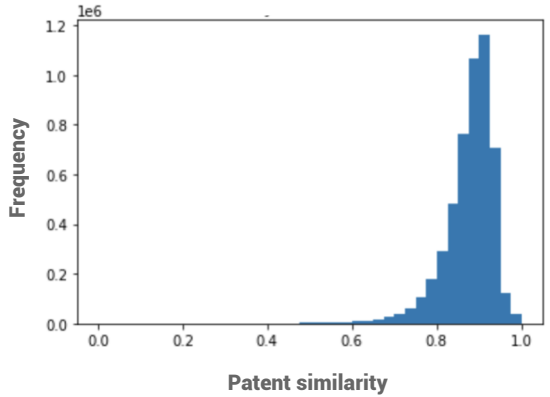
Top 10 words based on topic-word distribution for some topics

Topic Label	Top 10 Words
automobile	engine, fuel, cylinder, exhaust, combustion, injection, intake, internal, injector, inject
organic chemistry	composition, polymer, compound, comprise, organic, mixture, acid, agent, salt, weight
communication	partially, cable, linear, plug, comprise, fully, fiber optic, include, oscillate, ferrule

Patent Similarity

- Similarity between two patents using POS tags of their text
- Calculated average similarity between a patent and its backward cited patents

Patent similarity with backward citations



4 Approaches

1. Regression + Binning

- Regression models built using the full sample
- Number of citations is the target variable
- Obtained predictions are converted to categories
- Models considered:
    - Linear Regression
    - Poisson regression
    - Decision Tree
    - Random Forest

4 Approaches

2. Classification

- **Classification models built using the full sample**
- **Patent category is the target variable**
- **Models considered:**
  - **Logistic regression**
  - **Support vector machines**
  - **Decision tree**
  - **Random forest**

4 Approaches

3. Two-phase Classification

- **Phase 1**
  - **Classify patents into category 0 and non-zero**
- **Phase 2**
  - **Further classify non-zero category patents into categories 1 and 2**
- **Combine predictions from both phases into a final prediction**
- **Models considered in each phase:**
  - **Logistic regression**
  - **Decision tree classifier**
  - **Random forest classifier**

4 Approaches

4. BERT

- **Fine-tune BERT using independent claims in the mini sample**
- **Extract fine-tuned embeddings from the BERT model**
- **Use embeddings to classify patents into their categories**
- **Promising results obtained using the mini sample**
- **Training on full sample was computationally infeasible**

Results

Table contains the results corresponding to the best model of each approach

Metric	Approach		
	Classification	Regression	Two-phase Classification
	Decision Tree	Random Forest	Random Forest + Random Forest
Accuracy	45%	46%	46%
Macro Average F1 score	0.43	0.38	0.44

- **Accuracy - Percentage of correctly predicted samples**
- **Macro Average F1 score**
  - **F1 score conveys the trade-off between precision and recall for a particular category**
  - **Macro average F1 score - equally weighted average F1 score over all categories**

Results

Table contains the F1 Scores corresponding to the best model of each approach

F1 score of category	Approach		
	Classification	Regression	Two-phase Classification
	Decision Tree	Random Forest	Random Forest + Random Forest
0	0.55	0.11	0.40
1	0.34	0.60	0.52
2	0.40	0.36	0.41

- **Prefer a model with good F1 score of category 2 because:**
  - **Represents patents which receive 5 or more citations**
  - **These patents are of greater value**

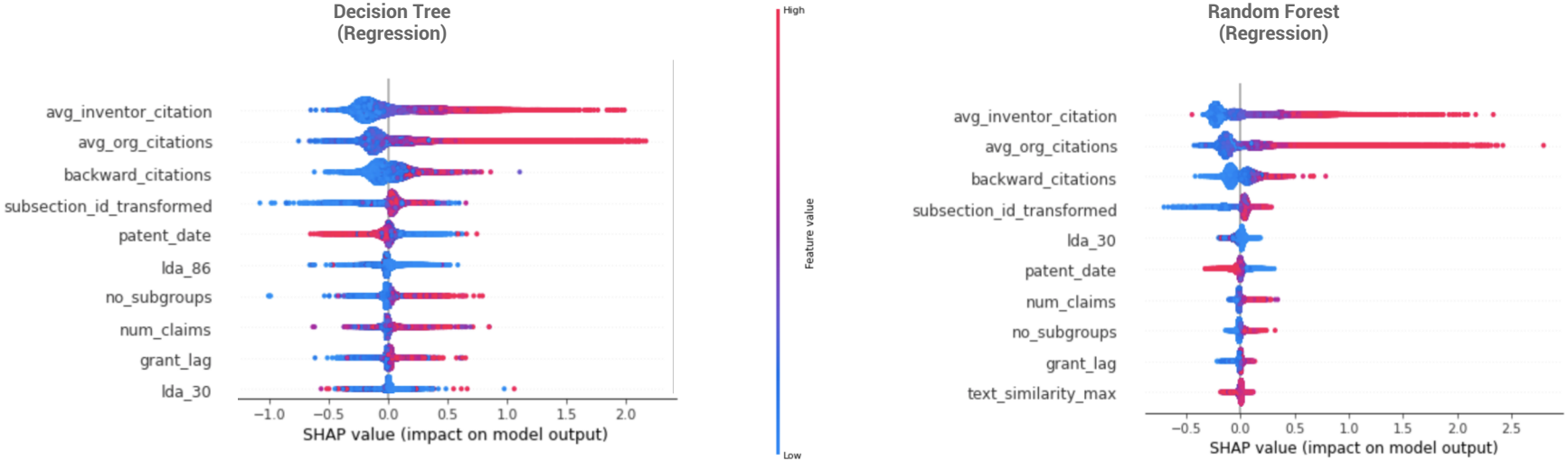
Interpretability

How to read the plot?

- The more a feature extends away from the 0 line, the more impact it has
- The direction in which it is (+ve or -ve) shows how it impacts the outcome
- The pink color shows higher value of the feature, and the blue lower

Topic modelling features:

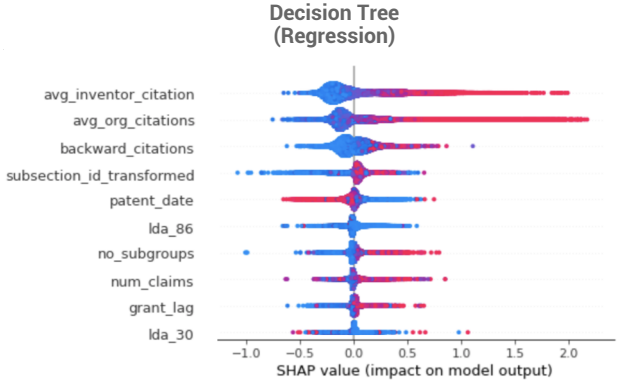
- lda_86 - topic related to flow and has words like spray, discharge, nozzle, liquid, supply etc.
- lda_30 - topic related to chemistry and has words like solution, coating, mixture, chemical etc.

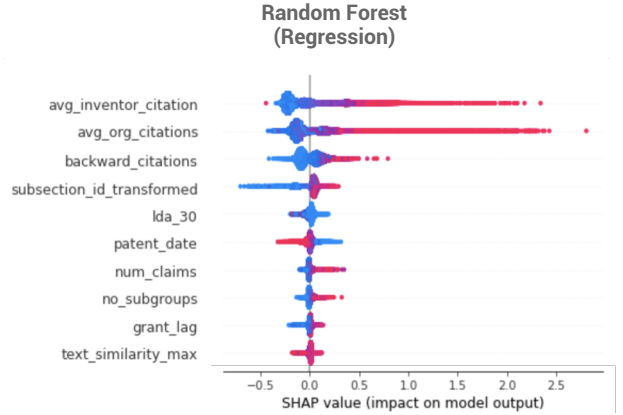


Interpretability

Summary

- **Model interpretations for decision tree and random forest, in regression approach**
- **Feature importance was similar in all the models, with slight ordering difference**
- **avg_inventor_citation, avg_organisation_citations, backward_citations seem to be the most important features - they positively impact forward citations**





Conclusion

- **Results suggest this is a hard problem for supervised models because:**
  - **Some latent factors are not captured in data**
  - **Models lack an extensive natural language component**
- **Two-phase approach performs best for our use case**
- **Features that contain prior information are most important:**
  - **inventor impact**
  - **organization impact**
  - **sub-section impact**

Future Work

- Explore Pegasus, a BERT model fine-tuned on a corpus of patents
- Incorporate time series aspect using ARIMA models
- Extract domain-specific terminologies from patents
- Leverage the entire text of patents
- Build ordinal regression models
- Scrape information of research papers that cite patents in our data

Thank you !