Predicting the Impact of a Scientific Paper

Vineet Malik, Aarushi Agrawal, Gautam Choudhary, Shreya Nasa*
malik83@purdue.edu,agraw218@purdue.edu,gchoudha@purdue.edu,snasa@purdue.edu
Purdue University
West Lafayette, IN, USA

Figure 1: Word cloud of various topics in papers published in 2017 in peer-reviewed conferences in Computer Science.

ABSTRACT

Authors are constantly unsure of whether their research will make an impact on the relevant fields and motivate researchers to draw inspiration from their work. They generally looks towards enhancing metrics like the citations, *h*-index, etc. in order to gain popularity amongst peers and scientific community. Towards this end, modeling and predicting the scientific impact of a research work is an important task. In this work, we model the impact of scientific articles in terms of citation counts both as a regression and classification problem. Subsequently, we survey and assess the performance of various data mining and machine learning approaches to the prediction task. We also analyze the impact of textual information, apart from the metadata, towards the model performance and find that meta data alone is a good indicator of citation counts.

KEYWORDS

Data mining, Scientific impact assessment, Citation prediction, Academic Social Networks, Machine Learning, Regression, Classification, Neural networks

1 INTRODUCTION

Over the past few decades, there has been a steady rise in the number of papers published making it impossible to read through all of these to understand their scientific impact. By scientific impact, we refer to an estimate for the quality of scientific papers and its influence on related research over time. Understanding the scientific impact and current trends in a scholarly environment is important for researchers to guide them on what to study and investigate, for viewers to identify important works, for recruiters to determine research potential of new hires, etc. Moreover, it is crucial for the authors as well to know the potential of their work to decide on what is the latest trend in their field and to restructure their goals with recent advances. This, along with the fact that knowledge of recent works is accessible with a single click, has made the world of research more competitive than ever where researchers are fighting for grants and approvals.

Prior works in this direction, attempt to quantify the "impact" of scholarly articles, authors, venues, etc., through various metrics like citations, *h*-index and impact factor, respectively. However, it is difficult to assess the quality of a research work solely based on these objective factors due to their subjectivity. This propels us to come up with a more effective way to evaluate the significance of any academic work due to the growing multitude of resources of scientific material that is made available to a broad audience.

One of the factors that play a major role in determining the success of a researcher is the citation count. Not only does it impact the recognition of its authors, but it is also an important factor that influences its retrieval by search engines. This calls for a model which could predict the future impact of a publication based on the paper and author information.

Solving this problem is a complex task as the number of citations varies over the lifetime of the paper and one needs to pay attention to this temporal aspect when designing the solution to the problem of predicting citations for any scientific paper. While some papers gain popularity in the early years, others might have a "cold" start and might not gain recognition in the first few years. The dynamic nature and skewed distributions (as shown later) makes is even more challenging. Also, the scale of the dataset is very huge (in Terabytes) which add to the complexity of the problem.

Existing works propose to use various machine learning approaches to capture the complex structure of the paper and authors, and predict the success of the paper in terms of citations. Through this project, we gain a deeper understanding of these models and the underlying parameters influencing the quality of a research

*All authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

paper. We work on various models to quantify the scientific impact of a paper by taking the citation count as a good proxy and use it as the ground truth for our models. We implement the solution using regression models to predict the exact citation count of a paper and using classification models by appropriately classifying the various works into corresponding buckets to capture its impact.

2 PROBLEM STATEMENT

We aim to tackle the following problems through this project:

(1) Predict the scientific impact of a paper in terms of its citation count
(2) Review and assess the importance of features used in the current literature
(3) Compare the impact of metadata features and the textual features on output prediction

3 LITERATURE REVIEW

Scholarly datasets (e.g., academic social networks) are a rich source of information and insights. There is vast ongoing research on efficient techniques for recommendation of related papers [12] and citations [7]. Similarly, identifying *rising* researchers in particular domains [5], mining trends in popularity of venues [19], ranking them [20], etc., are also broad topics of academic research.

We review the related works around the analysis of scholarly and bibliographic network datasets from various perspectives. We note that most of them revolve around proxy measures that determine the strength and quality of research articles like citation count, *h*-index, venue, etc. Citations received on an article tend to follow log-normal form [6, 18] and the skewed distributions make the regression task even more challenging. Authors of [15] analyse the relationship between the author's *reputation* and the citation counts of their papers. Along similar lines, work by [1] predicts *h*-index of an author based on their current academic journey (number of papers co-authored, time since first published paper, citation counts, etc.). Many works focus on predicting the citation counts based on their historical performance [3, 18] and meta-data (co-authors, cited articles) over a period of time. Some also account for their peer-review [13] to extract better indicators for their citation score. A recent work [8] lays foundational work for predicting citation potential for upcoming research papers which are yet to be published.

4 DATASET

The dataset that we work with comes from OpenAlex [1] which is open source and predecessor of the famous Microsoft Academic Graph (MAG) dataset. It is a large catalog of scholarly articles and their corresponding metadata accounting for around 1.6TB of storage when decompressed format. We make a note of other related datasets that we came across with, for instance MAG [2], AMiner [9], arXiv [4], PeerRead [10] but chose to work OpenAlex given it is open source API access and large repository of articles in various domains.

[1] OpenAlex: https://openalex.org/

4.1 Entities

The data is split into five basic entities:

(1) **Work**: This entity refers to the research articles, datasets, books, etc. that are produced by fellow researchers. A work can cite other works and contains information like paper title, paper abstract, bibliography information, fulltext, and its related meta data information like authors, date published, journal in which it got published, etc. Their are around 239M works till date in the dataset.
(2) **Author**: This refers to the authors who publish works and foster innovation. It contains information like name, affiliation(s), count of their works and citations, etc. There are around 213M authors identified in the dataset.
(3) **Venues**: The works are published or hosted at venues such as journals, conferences, workshops, etc. It contains demographics of the place and information like, identifiers, URL, count of works published and cited, etc. There are around 124k venues in the dataset.
(4) **Institutions**: Authors are affiliated with institutions such as a university or a company where they work at. There are around 110k institutions in the dataset.
(5) **Concepts**: Similar to WikiData, these can be thougth as semantic topics that works can be associated with. Some examples include computer science, phsysics, mathematics, etc. There are around 65k concepts in the dataset.

The Figure 2 show a visual representation of how the entities in the dataset interact with each other.

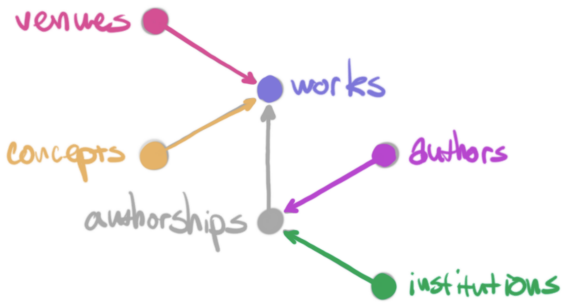

Figure 2: Entity graph of OpenAlex dataset [Source].

4.2 Data Collection

With MAG's retirement in Jan, 2022, OpenAlex was introduced as a replacement and also has a backward compatibility with it. The REST API provides a simple interface for fetching and retrieving chunks of data. Since the full dataset is large enough in storage, we chose to subset the data along some dimensions in order to get a manageable data to work with given the time and resource constraints. We collected the data of works published in 2017 in the field of computer science and in conference proceedings. Some basic filters were applied at this stage such as works having an abstract, references and that are not paratext, in order to obtain a more uniform structure. Along with this, the metadata of all the related entities including authors, venues, institutions and concepts were collected and parsed accordingly. The data was preprocessed by

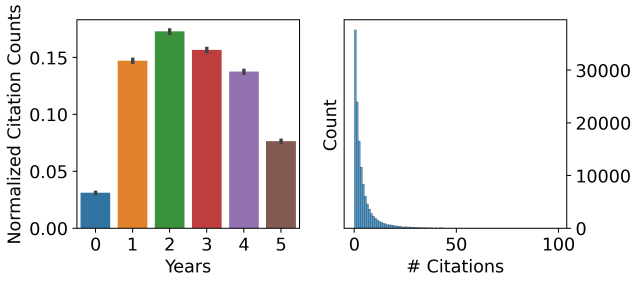

Figure 3: Normalized Citation Counts as a variation of time.

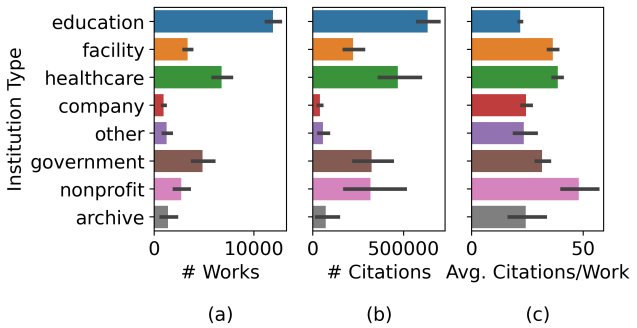

Figure 4: Counts of (a) works, (b) citations and (c) average citations per work of institutions across different sectors.

some heuristics like removing duplicate entries, records containing missing information, works where citation history exists prior to publishing, etc. leading to a more consistent nature of the data. Some of the descriptive statistics and raw-features of the processed data are shown in Table 1.

Table 1: Raw data extracted from different entities.

Entity	Counts	Raw data
Work	135k+	Abstract, Open accessibility, etc.
Authors	314k+	Mean citations count, works count
Insts	13k+	Mean citations counts, works count
Venues	1.7k+	Past citations counts, works count
Concepts	19k+	Titles, relevance scores

4.3 Exploratory Data Analysis

In this section, we describe the exploratory data analysis (EDA) performed on the processed data. While this is performed in conjunction with preprocessing by guiding our choices taken, this step is also helpful in understanding the overall nature of data. Apart from basic statistics, we seek to find some insights in this process. The proportion of journal articles seemed to be very high compared to peer-reviewed conference papers and thus, we confine ourselves with works published in conference proceedings as their citation life-cycles might be different in terms of duration and perhaps, counts.

Target Variable. For conference proceeding papers published in 2017, Figure 3(a) shows the log-normal [6, 18] trend of average citations garnered by a paper over the years post publishing in a venue. The maxima for the given data suggests that it takes around 2 years on an average for a paper to accumulate maximum citations within its lifespan of citations. This guides our choice for choosing the target variable as the cumulative sum of citations from 2017 to 2019. The distribution of target (regressor) variable y is reported in Figure 3(b) where the plot is filtered for works having less than 100 citations to highlight the skewness in data. Additionally, we normalize those counts by the months elapsed to account for papers who are published by year end than published in start of that year.

Institutions. While working the data, we analyzed top institutions based on the works they produced, citations gathered and average citations per work and plot their distribution categorized by differnt sectors. We found out that academic sector produced high amount of works followed by heath sector and industrial sector (for instance, companies) the least. However, the non-profit sector seems to produce more high quality work, at least in terms of citations, as shown in (c).

Concepts. Each work in the data is also associated with different number of *concepts* or fields of study along with a score signifying the strength of this association. Each of these are hierarchically related to each other at different levels, 'Computer-Science' being the high level concept at level 0. Figure 1 shows a wordcloud of these concepts and highlight top concepts.

5 FEATURE SELECTION

Given the above collected and processed data, we obtain the following set of features for later experiments. We categorize them into two high level buckets: (i) metadata features of each of the entities such as number of authors, open-access status, prior citation trends of venues, authors, etc., and (ii) textual features obtained through embedding texts such as abstracts, concepts, etc.

5.1 Metadata Features

(1) *Number of authors*: The count of authors of a paper was reported as one of the top features by [17].
(2) *Number of referenced works*: Papers that conduct surveys or literary reviews frequently cite more sources. Additionally, review papers are typically cited more frequently than other types of research.
(3) *Open accessibility*: Scholarly literature with open access is available for free and frequently has fewer onerous copyright and licensing restrictions than works with traditional publishing. So, open access papers have an edge in citation counts.
(4) *Publication month*: Although it may seem like an unneeded feature, certain studies [17] have shown that publications published early in the month tend to have higher citations.
(5) *Author's popularity*: Authors with a high reputation are frequently cited because they are read by many other researchers. The number of the author's publications and sum of all of their citation counts serve as our two benchmarks for popularity.

(6) *Publisher*: Publisher related parameters, such as the publication venue and publisher of the journal/conference, are the first characteristics thought to be indicative of a paper's likelihood to receive citations.
(7) *Institutions*: Similar to author attributes, we utilize the number of publications and the sum of their citations to assess the prominence of an institute.
(8) *Number of pages*: Number of pages in the work. This value was populated for less the 1% of the entries and was therefore ignored as a feature.

5.2 Textual Features

Feature vectors are obtained for abstract and concepts as per the following methods. These are then concatenated to make a 512-dimensional feature vector. Considering that we have a few metadata features, we perform PCA to find k most significant components and use the elbow method to find the optimal number of content features (16 features).

(1) *Abstract Embedding:* To capture the contextual information of the abstract, a Doc2Vec [11] model (gensim) is trained with 128-feature vectors and used to obtain respective embedding vectors for each work.
(2) *Concept feature vector:* Model embeddings ('all-MiniLM-L6-v2' sentence transformer model) are computed for each concept found in all the works and the concept feature vector for each paper is computed as a weighted sum based on the contribution of each of these concepts in the paper.

We predict Normalized Citation Count for a work given the metadata features and textual features. In order to compare the citation count for papers published at different points of time, we try two things. First, we consider all papers published in the same year. Next to improve the accuracy we normalize this over time using the number of months past its publication. So, let the citation count of a paper be CC_{paper}, and T_{paper} denote the number of months since the publication of the paper, then the normalized citation count is given by $\frac{CC_{paper}}{T_{paper}}$.

The first part of our work focuses on identifying the key features to predict the quality of a scientific paper. We extract the features for our models from the data and choose the most important features for computing the strength of the scientific paper. We infer the importance of features based on the model parameters (e.g., weights in linear regression model) and its correlation with the output strength value.

5.3 Correlation Study

In order to understand the significance of the various features used in the model, we computed their correlation with the citation counts.

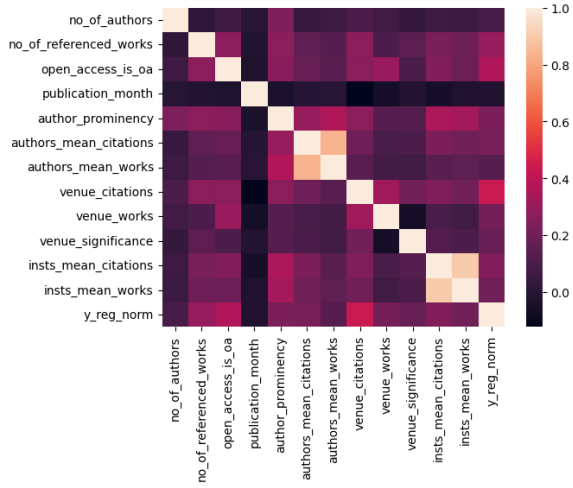

Figure 5: Correlation with metadata features.

As we can see from the heat map above, the citation count is highly correlated with average number of venue citations, the number of referenced works and the open access status of the work.

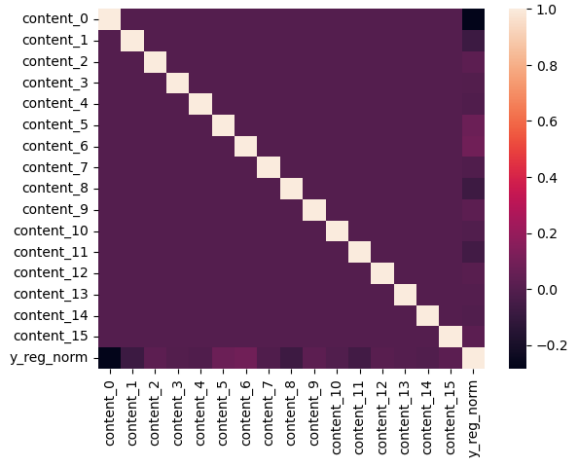

Figure 6: Correlation with textual features.

We used doc2vec model (gensim) on the paper abstract to compute a 128-feature vector representing the abstract of each paper and a model embedding ('all-MiniLM-L6-v2' sentence transformer model) to compute a 384-feature vector representing each concept. These concept embeddings were then weighted based on the contribution of these concepts in each paper and a corresponding 384-feature vector was computed for each work representing all the concepts in a paper. We thus computed a 512-feature vector representing the textual features of each paper by concatenating the abstract features and the concepts features. Since the number of metadata features was very small in comparison, we used PCA (Principal Components Analysis) to find the k most significant components in this data and used the elbow method to find the optimal number of content features (16 features for the given dataset).

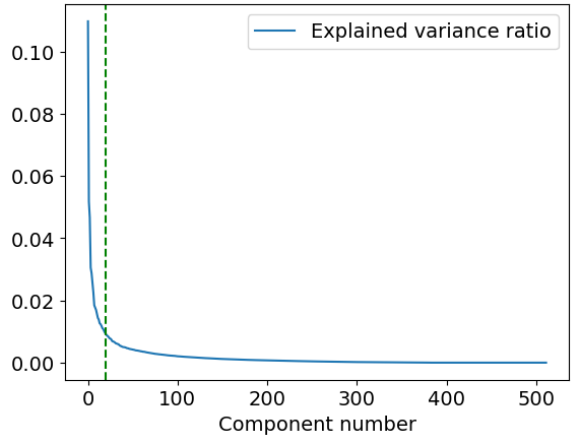

Figure 7: PCA explained variance ratio for texual features.

5.4 Causal Analysis

Although there may be a correlation between two features, this does not necessarily imply that the change in one feature is what led to the change in the values of the other feature. According to the concept of causation, there is a causal connection between the two events, meaning that one event results from the occurrence of the other event.

We examine the causal relationships among the features shown in Figure-8. These variables were chosen as they had strong correlation with the citation count. The directed acyclic graph in Figure-8 with features as nodes of the graph shows the causal relationships among the features. A directed edge from feature X to feature Y represents that X is a direct cause for Y.

We used the well-known PC [14] algorithm for inferring causal structure from data. The algorithm starts with a fully linked network and does conditional independence tests with increasing conditional set sizes to delete edges. For instance, the algorithm checks the conditional independence of X and Y for an edge between X and Y, and then it tests the conditional independence of X and Y given S with $|S| = 1$ to $|S| = N$, where N is the number of nodes that are neighbours of X or Y. If X and Y are independent given some set S, the edge between them is eliminated. The algorithm then orients the edges in the learned undirected graph in accordance with a predetermined set of rules. For all these conditional independence tests, we used the correlation coefficient or partial correlation coefficient under Fisher's z- transformation as the test statistic with the standard significance level of 0.05.

There is no outgoing edge from the citation count feature in the resulting causal structure depicted in Figure-8. Every other feature has a direct edge to the citation count as well. This supports the project's objective, which is to build a model to predict the citation count using metadata attributes.

6 MODELING

We measure the impact of a scholarly work using the metadata features. We model the problem in two ways: one as a regression task where we model a regressor to predict the citation total number of citations and the other as a classification task where a paper is classified into different classes based on the citation count.

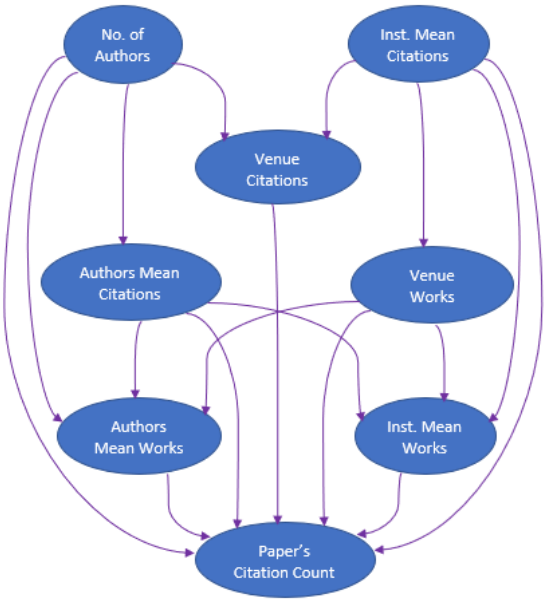

Figure 8: Causal graph on features.

6.1 Regression

For the regression task, we test the performance on different regression models, namely Linear Regression, lasso regression, Gradient Boosting regression, SGD Regressor, XGB Regressor, MLP Regressor, and Zero Inflated Regressor. The regression performance is measured with textual features and metadata features together and separately.

6.2 Classification

For the classification task, we classify works in different buckets. For this, we cluster the citation counts for data using Gaussian Mixture Model (GMM) [16] and evaluate the quality of clusters using Silhoutte score and BIC Score. BIC score measures the goodness of GMM in predicting the clusters while penalizing large number of clusters. Silhouette score is used to evaluate the performance of any clustering technique by measuring the intra-cluster and inter-cluster distance. A higher Silhouette score means better clustering. Figure shows the graph for BIC Score and Silhoutte Score for varying number of clusters. We choose the number of clusters with the highest gradient in BIC score, and classify the works into 2 clusters thresholded at citation count 5. Works with a citation count of 6 or more are labeled influential works(class 1). And works with 5 or fewer citations are labeled less influential works(class 0).

We classify the works using logistic regression, XGBoost Classifier, Decision Tree Classifier, K-Nearest Neighbor Classifier, Gaussian Naive Bayes Classifier, Support Vector Classification(SVC),and Multi-layer Perceptron Classifier(MLPClassifier). The classification performance is measured with textual features and metadata features together and separately on all the classifier algorithms.

Table 2: Regression Results.

Model	Train			Test		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Zero Regressor	219.457	179.765	-2.039	276.760	199.181	-1.074
Mean Regressor	125.859	103.032	0.000	192.987	124.299	-0.009
With Metadata Features						
Linear Regression	**93.396**	**75.216**	**0.450**	**162.252**	**104.419**	**0.287**
Lasso Regression	98.501	79.195	0.388	171.320	110.516	0.205
Gradient Boosting Regressor	36.944	27.676	0.914	189.469	113.778	0.028
SGD Regressor	98.938	76.434	0.382	176.920	110.523	0.152
XGB Regressor	20.937	12.266	0.972	208.344	127.052	-0.176
MLP Regressor	142.223	107.545	-0.277	234.734	153.354	-0.492
Zero Inflated Regressor	**93.396**	**75.216**	**0.450**	**162.252**	**104.419**	**0.287**
With Textual Features						
Linear Regression	**108.208**	**87.890**	**0.261**	**169.449**	**115.236**	**0.222**
Lasso Regression	124.910	102.066	0.015	191.946	123.298	0.002
Gradient Boosting Regressor	34.239	24.792	0.926	201.541	119.286	-0.100
SGD Regressor	156.188	116.254	-0.539	203.007	124.628	-0.116
XGB Regressor	17.285	8.883	0.981	218.719	130.758	-0.296
MLP Regressor	176.481	138.001	-0.966	252.467	173.205	-0.726
Zero Inflated Regressor	**108.208**	**87.890**	**0.261**	**169.449**	**115.236**	**0.222**
With Metadata and Textual Features						
Linear Regression	**86.911**	**70.681**	**0.523**	**161.035**	**104.127**	**0.298**
Lasso Regression	98.501	79.195	0.388	171.320	110.516	0.205
Gradient Boosting Regressor	28.716	20.446	0.948	216.795	119.530	-0.273
SGD Regressor	96.734	75.185	0.409	168.938	105.424	0.227
XGB Regressor	14.526	7.013	0.987	215.743	124.371	-0.261
MLP Regressor	108.945	81.426	0.251	227.375	146.444	-0.400
Zero Inflated Regressor	**86.911**	**70.681**	**0.523**	**161.035**	**104.127**	**0.298**

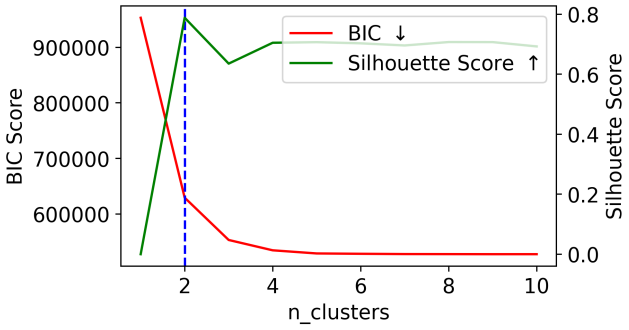

Figure 9: BIC Score and Silhouette Score for GMM on citation count.

7 EVALUATION

Estimating the impact of a paper can be expressed as a regression problem that predicts a numerical value, such as the number of citations, or as a classification problem that classifies the work as influential or not based on the number of citations.

We have sampled the input data to balance the dataset as much as possible since most of the papers had very low citation counts. About 43% of the papers had zero citations and about 55% had between 1 to 4 citations. To tackle this imbalance, we randomly sampled a fraction of the papers from these two bins. The dataset was further split into the training and test datasets. The training dataset consists of 11920 entries and test dataset consists of 2980 entries. Considering the computationally expensive task of training language modes for abstract, we decided to not use K Fold cross-validation for evaluating model performance.

In order to evaluate the performance of our model, we compare the output of the various regression models with the citation count of the scientific paper. We have used the root mean square error (RMSE), mean absolute error (MAE) and R^2 score to measure the error in the prediction of citation counts and compare the performance of various models. We have used the accuracy, F1-score and AUC values for the classification problem of identifying whether the paper is influential or not - classified based on the reasoning described in the Section 6.2. All these scores are weighted (based on the count of papers in various bins or classes) when possible (KNN, for example, does not support sample weights) to ensure that the models aren't biased to predict lower citation counts for all works.

Table 3: Classification Results.

Model	Train			Test		
	Accuracy	F1	AUC	Accuracy	F1	AUC
Majority Classifier	0.980	0.084	0.500	0.978	0.087	0.500
Stratified Classifier	0.967	0.115	0.500	0.974	0.121	0.501
Random Classifier	0.443	0.538	0.500	0.650	0.549	0.500
With Metadata Features						
LogisticRegression	0.980	0.084	0.839	0.978	0.087	0.854
XGBoost	0.980	0.084	0.688	0.978	0.087	0.681
KNeighborsClassifier	0.865	0.873	0.933	0.890	0.834	0.813
GaussianNB	**0.911**	**0.808**	**0.811**	**0.944**	**0.815**	**0.827**
DecisionTreeClassifier	1.000	1.000	1.000	0.874	0.794	0.711
SVC	0.980	0.084	0.347	0.978	0.087	0.358
MLPClassifier	**0.897**	**0.863**	**0.884**	**0.929**	**0.859**	**0.884**
With Textual Features						
LogisticRegression	0.980	0.084	0.712	0.978	0.087	0.732
XGBoost	0.980	0.084	0.500	0.978	0.087	0.500
KNeighborsClassifier	**0.678**	**0.822**	**0.868**	**0.640**	**0.743**	**0.707**
GaussianNB	**0.930**	**0.594**	**0.711**	**0.947**	**0.594**	**0.724**
DecisionTreeClassifier	1.000	1.000	1.000	0.613	0.705	0.583
SVC	0.980	0.084	0.355	0.978	0.087	0.344
MLPClassifier	0.528	0.787	0.827	0.617	0.765	0.763
With Metadata and Textual Features						
LogisticRegression	0.980	0.117	0.859	0.978	0.107	0.875
XGBoost	0.980	0.084	0.688	0.978	0.087	0.681
KNeighborsClassifier	0.903	0.882	0.938	0.927	0.845	0.836
GaussianNB	**0.895**	**0.822**	**0.827**	**0.933**	**0.825**	**0.844**
DecisionTreeClassifier	1.000	1.000	1.000	0.873	0.790	0.708
SVC	0.980	0.084	0.510	0.978	0.087	0.506
MLPClassifier	**0.930**	**0.895**	**0.942**	**0.941**	**0.869**	**0.884**

7.1 Results

We have used two different ways to measure the impact of a scholarly work using metadata features and textual features. Overall we see that in both approaches, the model gives better performance with only metadata features as compared to only textual features. We noticed that the models can give a good accuracy (when we do not use any sample weights or class weights) even while predicting all zeros (majority regressor model) since a significant number of works have 0 citation count. To counter this, we use Zero Inflated Regressor (uses classification as first step to predict whether the output as zero and then uses regression to predict the number of citations). The results for various regression and classification models are shown in Table 2 and Table 3.

7.2 Comparing Metadata and Textual Features

To understand the significance of textual features of the paper on predicting its scientific impact, we trained the various models using the metadata features only, the textual features only and all the features combined. As evident from the results tables below, the models trained with only metadata features generally perform better than the models trained with only content features. The models trained with all the features combined tend to perform considerably better than the individual models which leads us to conclude that the content features play a significant role in determining the impact of a scientific paper but these need to be analyzed and mined properly to have a substantial impact on the models.

7.3 Insights

When modeled as a regression task to predict the normalized citation count, we see that the prediction performance improves for SGD Regressor, Linear Regressor, and Zero Inflated Regressor when both textual and metadata features are used. For Gradient Boosting Regressor and XGB Regressor, we see small training errors but the models perform poorly on test data due to overfitting. For the Lasso Regression model, the textual features do not change the model performance. When modeled as a classification task, the classification score improves for Logistic Regression, KNN Classifier, Gaussian Naive Bayes Classifier and SVC with additional textual features as compared to only using metadata features. This could be because

the textual features do not fully represent the content and innovations made by the paper. Also, a dataset of only 12k abstracts is not enough to learn a language model. We see the impact of unbalanced datasets (output is biased towards low citation counts) in the results as the mean regressor performs considerably well and even outperforms a few other trained regression models. For the regression models, the linear regression and zero inflated regression models perform better than the other regression models implying that the regression models trained are properly penalized for predicting low citation counts, consequently they perform equally well for higher citation counts as well. For the classification problems, we see that Gaussian Naive Bayes consistently outperforms the other models for all combinations of features, which leads us to believe that the Naive Bayes assumption holds for the features we have used in our models, i.e, they are conditionally independent of each other. Logistic regression, on the other hand, has a very high accuracy but a very low F1 score, again pointing towards the imbalance in dataset and biased model training even after adding sample and class weights. KNN and MLP classifier also perform considerably well even though they do not support weighted samples in loss function, hence the imbalance in the prediction counts has a lower impact for the classification task as compared to the regression task.

8 FUTURE WORK

The problem of the influence of a scholarly paper is very complex. In this project, we have studied this as a classification problem and regression problem. Some papers can have a cold start and might get the academic community's attention in later years. While this would be an impossible task to predict, a simpler question could be to evaluate the influence in the near future considering the temporal aspect of input features. Another problem that has often perplexed researchers is the venue they can submit their paper to. Since conferences don't release data on rejected submissions, this would be modeled as a Positive Unlabeled Learning task. In this project we have only extracted features from concepts and abstract which do not measure the significance of the contribution made. Extracting features on other sections of the work and using them in the model could give better results. While the impact is measured in terms of citation count here, there is space to explore other metrics as well that we can used to quantify the impact.

9 CONCLUSION

Feature engineering for prediction of the scientific impact of a paper is a very intricate and challenging task as there is a plethora of features available that may contribute to the citation count of a paper and its scientific importance in general. This may not be limited to the metadata and content features as considered in our project, but also to some unpredictable external factors like the time at which the paper is published. If the paper is published when the particular topic is receiving great attention in academia and industry, it is likely to get higher citation counts. Without taking this external unpredictable factors into account, our models focus on the easily available features related to the paper when it is published and predict the citation count with a high accuracy. We tackle the problem of imbalanced datasets in this project and thus train the models and compute the metrics such that the loss function gives poor results for biased models. The classification models perform a little better than the regression models as the imbalance in the citation counts in datasets is minimized when we bucket them into two classes. The improvement in performance metrics with the inclusion of textual features indicates that in an academic setting, the content of the paper matters and it's not just the metadata, which contains the attributes of the authors, institutions and venues, that plays an important role. The major contribution of any work is explained in the methodology and the results sections of the work, which have not been taken into consideration in our project. We can confidently say that the project demonstrates well that the influence that a work has in the scientific community is neither completely unpredictable nor can it be entirely known, but its scientific impact can be predicted with a high confidence and consequently used for various purposes as described earlier.

10 INDIVIDUAL CONTRIBUTION

While all the team members contributed to all parts of the project, their major analysis and findings can be broadly categorized as follows:

(1) Gautam Chaudhary - Dataset exploration, exploratory data analysis
(2) Shreya Nasa - Feature extraction and understanding the impact of content features
(3) Aarushi Agrawal - Analyzing regression models
(4) Vineet Malik - Analyzing classification models

REFERENCES

[1] Daniel E Acuna, Stefano Allesina, and Konrad P Kording. 2012. Predicting scientific success. *Nature* 489, 7415 (2012), 201–202.
[2] Yang Song Hao Ma Darrin Eide Bo-June (Paul) Hsu Arnab Sinha, Zhihong Shen and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion))*. ACM New York, NY, USA, 243–246.
[3] Xiaomei Bai, Fuli Zhang, and Ivan Lee. 2019. Predicting the citations of scholarly paper. *Journal of Informetrics* 13, 1 (2019), 407–418.
[4] Bierbaum M. O'Keeffe K. P. Alemi A. A. Clement, C. B. (2019). On the Use of ArXiv as a Dataset. In *arXiv:1905.00075*.
[5] Ali Daud, Muhammad Ahmad, MSI Malik, and Dunren Che. 2015. Using machine learning techniques for rising star prediction in co-author network. *Scientometrics* 102, 2 (2015), 1687–1711.
[6] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2014. Will this paper increase your h-index. *Scientific Impact Prediction. ArXiv e-prints* 1412 (2014).
[7] Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C Lee Giles. 2014. Refseer: A citation recommendation system. In *IEEE/ACM joint conference on digital libraries*. IEEE, 371–374.
[8] Song Jiang, Bernard Koch, and Yizhou Sun. 2021. HINTS: citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *Proceedings of the Web Conference 2021*. 3158–3167.
[9] Limin Yao Juanzi Li Li Zhang Jie Tang, Jing Zhang and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*. 990–998.
[10] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635* (2018).
[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.
[12] Joonseok Lee, Kisung Lee, and Jennifer G Kim. 2013. Personalized academic research paper recommendation system. *arXiv preprint arXiv:1304.5457* (2013).
[13] Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. 2019. A neural citation count prediction model based on peer review text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 4914–4924.

[14] Christopher Meek. 2013. Causal inference and causal explanation with background knowledge. *arXiv preprint arXiv:1302.4972* (2013).

[15] Alexander Michael Petersen, Santo Fortunato, Raj K Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H Eugene Stanley, and Fabio Pammolli. 2014. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* 111, 43 (2014), 15316–15321.

[16] Douglas A Reynolds. 2009. Gaussian mixture models. *Encyclopedia of biometrics* 741, 659-663 (2009).

[17] Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics* 14, 3 (2020), 101039.

[18] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. 2016. On Modeling and Predicting Individual Paper Citation Count over Time.. In *Ijcai*. 2676–2682.

[19] Muhammad Azam Zia, Zhongbao Zhang, Guangda Li, Haseeb Ahmad, and Sen Su. 2017. Prediction of rising venues in citation networks. *Journal of advanced computational intelligence and intelligent informatics* 21, 4 (2017), 650–658.

[20] Muhammad Azam Zia, Zhongbao Zhang, Ximing Li, Haseeb Ahmad, and Sen Su. 2017. ComRank: joint weight technique for the identification of influential communities. *China Communications* 14, 4 (2017), 101–110.