Predicting Grant Success: A Multimethod Approach Using TDA, NLP, and Bibliometric Analysis

Blake Bates, Ian Luff, Eric Sung

December 17, 2024

1 Introduction

In this analysis, we aim to identify patterns in a dataset of historical grant proposals submitted to the Department of Defense by professors at the University of Arizona. Our objective is to apply tools learned in this class, along with additional methods, to uncover trends that could assist future students and professors in successfully applying for Department of Defense funding. The dataset was obtained through the National Security Initiatives at the University of Arizona.

The dataset contains approximately 1,600 rows and seven key columns. These columns include the proposal title, proposal status, college name, lead investigator, sponsor name, amount requested, and submission date. To maintain privacy, the names of the professors have been anonymized. One interesting approach we explore involves an application of topological data analysis (TDA). This idea was inspired by the work of Jerome Roehm [1]. In his work, TDA is used to measure the diversity of a basketball team. Instead of basketball teams, we apply this idea to academic units such as the College of Engineering, and instead of basketball players, we analyze the most active professors—those with the highest number of submissions. Our goal is to determine whether a diverse spread of proposal strategies leads to a higher success rate for approvals by the Department of Defense.

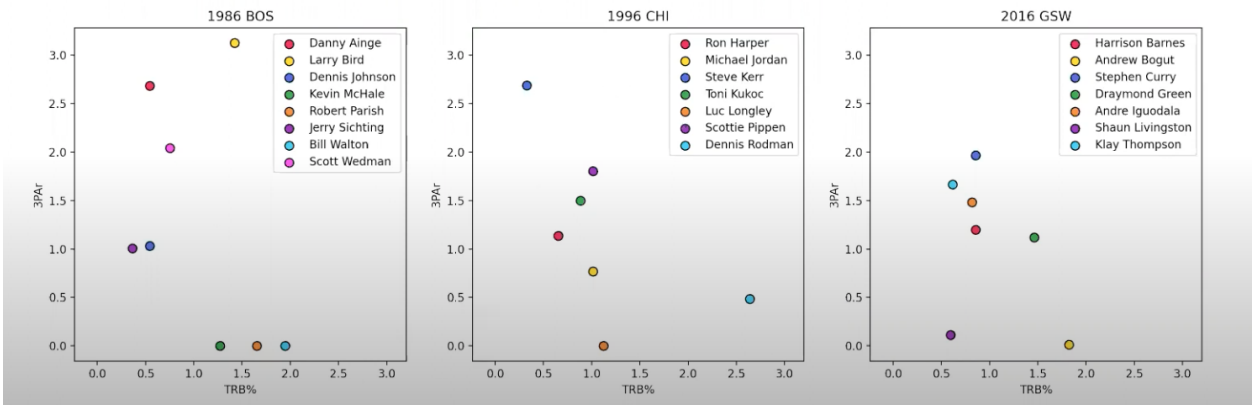
After completing this analysis, we gained a clearer understanding of what to prioritize in our remaining research. Our second approach involved a Natural Language Processing (NLP) model, which is primarily driven by the abstract titles, keywords, and author. Originally we tested the plausibility of this idea by just considering the title of the paper, which yielded a high success rate when considering just one feature but was not accurate enough. Thus, we decided to increase the number of features to three by also considering keywords and the lead investigator for each proposal. One hurdle was identifying keywords that correlate with a higher likelihood of securing funding from the DoD. Once we complied a satisfactory list of keywords, our model had a drastic increase in accuracy. These keywords and methodologies will be discussed later on in greater detail.

Our final approach analyzed the bibliometric metrics of professors to determine their correlation with grant approval success. We augmented the dataset with h-index, i10-index, and total citation counts, subdivided into cumulative and recent scores. Using the logistic regression, random forest, and gradient boosting models, we identified the i10-index and total citation count as the strongest predictors of success. Gradient boosting emerged as the most effective model, achieving the highest accuracy and AUC-ROC. These findings highlight the role of bibliometric data in understanding grant approval trends and will be discussed further in the following sections.

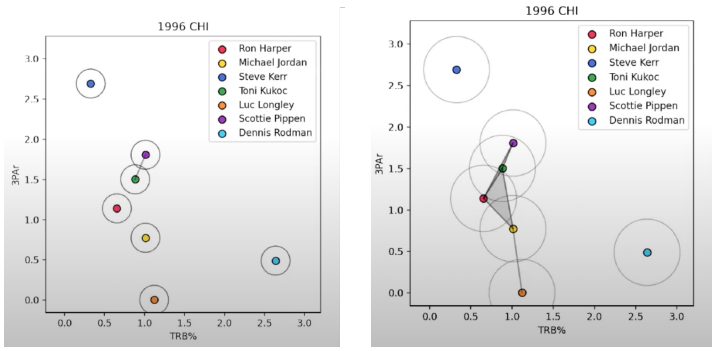
2 Topological Data Analysis

Topological Data Analysis (TDA) is a framework for analyzing the topology of datasets. It is particularly appealing because it is robust to noise, meaning that even if the data is slightly perturbed or contains noise, TDA tools can still identify the underlying topological structure. A central tool in TDA is Persistent Homology (PH), which examines the persistence of topological features as we filter the dataset. Features that persist across a wide range of the filtration are considered significant, while those that disappear quickly are treated as noise.

An interesting application of persistent homology (PH) is found in basketball. Jerome Roehm utilized topological data analysis (TDA) to measure the diversity in playstyle of a given basketball team. He achieved this by plotting data for specific teams during a particular year. In the figure below, the x-axis represents the team's average total rebound percentage for that year, while the y-axis represents the team's average number of three-point attempts per game.



A team can be considered more diverse if the datapoints are more spread out. Roehm applied TDA to quantify this diversity of playstyles. The method involves constructing a simplicial complex (essentially a type of graph) by growing "balls" around each datapoint and connecting them with edges whenever the balls intersect. This process is illustrated in the images below:



As the radii of the balls increase, the number of connected components decreases. TDA provides tools to track the lifespan of these connected components. The average lifespan of the connected components serves as a measure of diversity.

Roehm then investigated the correlation between this diversity measure and a team's overall success. His analysis revealed a relatively significant positive correlation, suggesting that greater diversity in playstyle is associated with better team performance.

For our project, we applied this idea to our dataset. Instead of a basketball team, we considered the different colleges at the University of Arizona, and instead of team players, we considered the top 5 most active professors submitting proposals under each given college. We wanted to answer the question: does having diverse proposal strategies result in more success in funding from the Department of Defense (DoD)?

We considered multiple combinations of choices for the x-axis and y-axis: number of DoD sponsors targeted vs. total submissions, acceptance rate vs. average grant amount requested, number of funded proposals per year vs. total proposal submissions per year, and acceptance rate vs. total submissions.

The colleges we analyzed were the College of Optical Sciences, the College of Aerospace and Mechanical Engineering, the College of Electrical and Computer Engineering, the College of Psychiatry, and the College of Materials Science and Engineering.

Using TDA to compute the diversity measure, we then computed the correlation. Here are the results:

Metric Pair	Correlation
number of DoD sponsors targeted vs. total submissions	-0.8415
acceptance rate vs. average grant amount requested	-0.1791
number of funded proposals per year vs. total proposal submissions per year	-0.8329
acceptance rate vs. total submissions	-0.8479

Table 1: Correlation Results

We see a strong negative correlation between the success rate of receiving funding from the DoD and the measure of diversity. This did not surprise us after some thought. It makes sense that the DoD has specific interests and therefore rewards specific strategies for receiving funding.

Our next idea was to focus on the key topics in the proposal titles, as the DoD is likely receiving a large number of proposals daily and filters through these proposals using key words or phrases.

In conclusion, the TDA approach did not give us the results we initially expected, but it guided our endeavor, allowing us to understand that the key to success in funding is likely a very specific proposal strategy, as opposed to a diverse proposal strategy.

3 Natural Language Processing

Due to the nature of our data, we needed a technique for processing raw text or converting the text into numerical values before processing. While discussing how to convert the text into numerical data, there were a lot of assumptions and uncertainties that ultimately drove us away from this approach and towards an NLP approach. The advantages of an NLP approach, beyond the nature of our data, is that this methodology allows for flexibility as the focus of the DoD changes.

As technology, the country, and the world change, the focus of the DoD also changes. As such, the types of research that the DoD chooses to fund and by how much also change. This fact insinuates the importance of titles and keywords that can capture the essence of the DoD's priorities. Furthermore, as funding is a form of investment, we find it reasonable to assume that the DoD will choose to invest in people with known capacities to have a return on investment over taking risks on people of little renown. These perspectives motivated our choice of the three features: title, keywords, and author.

For the purposes of this project, we decided to utilize Python over R, as python has more pre-built libraries, pre-trained models, and a larger community support for NLP codes. To be more specific with how we are processing our data, as NLP is a rather large category, our code utilizes a Transformer Neural Network. The purpose of using a Transformer Neural Network over any other Neural Network is that Transformers can capture context and relationships between words more efficiently than other Neural Networks. This is especially important when processing scientific articles, where titles can be relatively long with words that should be associated with one another far apart. A Transformer Neural Network will recognize these relationships. To further improve the accuracy of our Neural Network, we had the model pre-trained on scientific papers and articles. This allowed our Neural Network to associate technical language more efficiently, than if it were pre-trained on a model with literature or written media.

Once we laid this foundation and wrote our code, we had to prep our dataset. In its purest form, our data has up to seven features and it has four labels. The labels are either "accpeted", "rejected", "withdrawn", and "pending". The issue with the "withdrawn" and "pending" labels, is that it is hard to determine the cause for each lable. A project that is labeled as "pending" could be due to a back log at the DoD or it could be a secondary choice for funding. With the "withdrawn" label, this is typically done by the person submitting the proposal, but the reasons vary. It could be that they are no longer seeking funding form the DoD, or at all, or it could be because they plan to revise their proposal and submit it another time. Given the uncertainty around these labels, we chose to omit them for the time being. Thus, or problem has become a binary classification problem with three features. Therefore our outputs belong to the set $\{0, 1\}$, with 0 signifying a rejection and 1 signifying an acceptance.

To verify the functionality and efficacy of our code, we started off by only considering one feature: the title of the article. The rational behind this decision was to check if our model could process the technical language sufficiently and to check if our predictions belonged to the output set defined above. For this test,

we had python take in our data as an excel file and randomly choose data points such that 75% of the data was for the training data set and 25% was for the test data set. The choice of these percentages was to have a sufficiently sized training set to train our model, but to also be semi-accurate to the reality of future predictions. In the future, if we choose to use this model again, our training data will be several years worth of proposals and prediction just one year worth of proposals. This means that our training data set will be significantly larger than our test data set. Upon passing our data through our model, the model was able to correctly identify 214 out of 305 data points. This yields an accuracy of 70.16%. This result was very much unexpected and seemed too good to be true. To check our Neural Network and ensure that our results were semi-accurate, we decided to determine the training error. We had a our model predict our training data and it misclassified 78 data points out of 913, which yields an accuracy of 91.45%. This is indicative that our neural network is over fitted to our training data and is not well suited for new data. Regardless, we decided to see if our model was at least consistent across trials.

To ensure that our model was consistent across trials, we ran our code several times with the same 75% training data and 25% mixture, but this time we changed the random seed at each iteration. Using the random seeds of 42, 1, 10, and 21 we obtained the following results.

Trial, Random Seed	**Model Accuracy**
Trial 1, 42	70.16%
Trial 2, 1	69.84%
Trial 3, 10	67.54%
Trial 4, 21	74.10%

Table 2: Model Accuracy: Title Only

As we can see from the table, our model has an average accuracy of about 70.41%, which is very decent when considering only one feature. Furthermore, across all our trials our model had a prediction accuracy that was relatively close to our initial accuracy of 70.16%. Since the random seed changes which data belongs to the training set and which data belongs to the test set, these results suggest that our model is flexible with the training and test sets. With these results in hand, we decided to add our keyword and author features to see if our results improved. With some minor changes to our code, we added the extra features to our data and arrived at the following results.

Trial, Random Seed	**Model Accuracy**
Trial 1, 42	71.47%
Trial 2, 1	68.20%
Trial 3, 10	66.88%
Trial 4, 21	74.75%

Table 3: Model Accuracy: Title, Keywords, and Author

Based on the table above, the results are relatively non-conclusive. In two cases the accuracy increased, with the greatest increase being 1.31%. In the other two cases our accuracy decreased, with the greatest decrease in accuracy being 1.64%. These results are counterintuitive to what we were expecting. Considering the fact that we believe our model is overfitting to our training data, this would be the first place we would investigate moving forward on how to rectify this issue. Furthermore, this problem could occur because some data points have a higher correlation to their predictions than others. For example, articles with titles that are longer or more descriptive may yield better accuracy, compared to titles that are less descriptive. If our training data contains more of the less descriptive titles, then our model will have trouble generalizing to the test data. All of these issues would need to be looked at in greater detail in the future and rectified before utilizing this model in a more practical setting.

4 Predictive Modeling of Grant Outcomes Using Bibliometric Data

4.1 Bibliometrics: *h*-index and i10-index

In the modern academic landscape, the sheer number of researchers across all disciplines, coupled with an exponentially growing volume of published papers each year, can make it daunting for novice researchers to identify impactful work and its contributors. To address this challenge, several bibliometric metrics have been introduced to quantify a researcher's productivity and the influence of their work. These metrics primarily rely on the number of publications and the citations they accumulate. Among the most widely recognized metrics is the *h-index*, followed closely by the *i10-index*. It is well established that higher bibliometric scores, such as a high h-index, confer significant advantages, including increased acceptance of research fellowships, publications in prestigious journals, and recognition through awards [2].

For our project, we focus on three key bibliometric measures: the h-index, i10-index, and total citation count. To capture both long-term and recent research productivity, these metrics are further subdivided into two categories: cumulative scores (total) and scores from the last five years, resulting in six distinct parameters.

The h-index is defined as the largest number h such that h of a researcher's articles have been cited at least h times each [3]. Formally, if f represents the function mapping a researcher's publications to their respective citation counts, then the h-index can be expressed as:

$$h\text{-index } (f) = \max\{i \in \mathbb{N} : f(i) \geq i\}.$$

The i10-index, on the other hand, is defined as the number of publications that have received at least 10 citations [4]. This straightforward yet effective measure complements the h-index by highlighting the breadth of a researcher's influence across moderately cited papers.

4.2 Model Comparison

Our initial dataset consisted of 1678 entries, which required significant preprocessing. Since the provided dataset did not include bibliometric metrics, we gathered this data for the 470 unique authors. Google Scholar served as the primary source; for authors without a Google Scholar profile, we used Semantic Scholar, which only provides the h-index and total citation count. Entries for authors lacking bibliometric data were removed, reducing the dataset by 19 entries. To focus exclusively on grant outcomes categorized as *approved* or *rejected*, we removed entries with statuses such as *deactivated*, *withdrawn*, or *pending*, further reducing the dataset by 474 entries. This preprocessing resulted in a cleaned dataset comprising 1204 valid entries. Given that Semantic Scholar offers limited bibliometric metrics compared to the more detailed information from Google Scholar, we subdivided the dataset into Google Scholar and Semantic Scholar subsets. We opted to use only the Google Scholar subset, which accounted for 86% of the cleaned data and offered greater accuracy and detail, thereby minimizing the risk of bias while ensuring a robust analysis.

Due to the binary nature of this analysis, we first employed a logistic regression model as a baseline approach. The primary objective was not to maximize predictive accuracy but rather to probe the relationships between bibliometric features and grant approval outcomes. The predictors included in the model were the h-index, i10-index, and total citations. Logistic regression achieved a Test AUC-ROC of 0.570 and an Accuracy of 53.9 %, indicating poor predictive performance, only marginally better than random guessing. Feature importance analysis, based on the model's coefficients, revealed that the h-index contributed the most to predictions, followed by the i10-index, while total citations had minimal impact. Residual analysis further highlighted the model's limitations, with a bimodal distribution of residuals suggesting frequent underestimation and overestimation of probabilities. Additionally, significant overlap was observed in the predicted probabilities for accepted and rejected grants, indicating the model's inability to effectively separate the two classes. These findings underscored the need for more sophisticated, non-linear models to capture the complex relationships in the data.

Following the limitations observed in logistic regression, we implemented Random Forest, an ensemble learning method capable of handling non-linear relationships and feature interactions. The untuned Random Forest model, trained using the same three predictors (h-index, i10-index, and total citations), significantly improved predictive performance, achieving a Test AUC-ROC of 0.745 and an Accuracy of 68.4 %. Feature

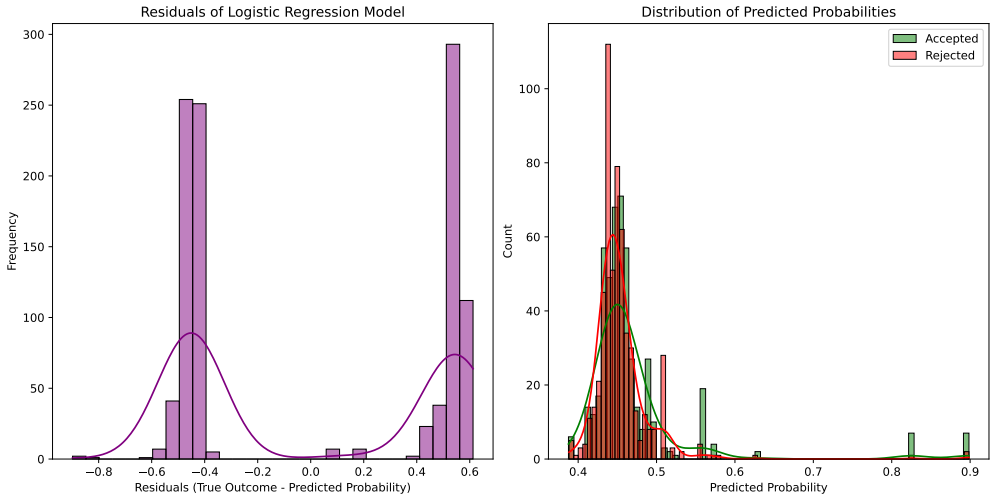


Figure 1: Residuals and Predicted Probability Distribution for Logistic Regression.

importance analysis revealed that total citations and i10-index were the most influential predictors, reinforcing their importance in determining grant approval, while the h-index contributed moderately.

To further illustrate the relationships between features and model predictions, we analyzed partial dependence plots (Figure 2), which highlight the marginal effect of each predictor on the predicted probability of grant approval. The i10-index exhibited a stable and positive association, while total citations displayed diminishing returns for higher values, suggesting saturation effects. The h-index, however, demonstrated a less clear and noisier relationship.

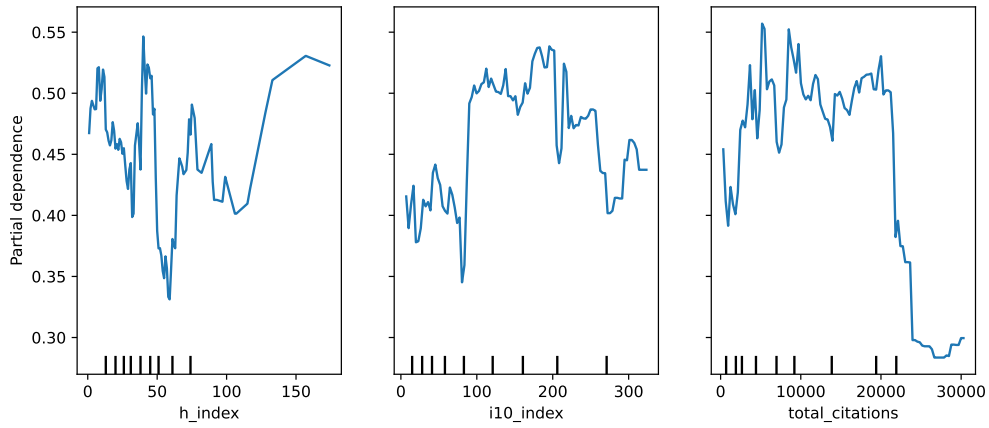


Figure 2: Partial Dependence Plot for Untuned Random Forest.

To further enhance model performance, we fine-tuned the Random Forest using grid search to optimize key hyperparameters, such as the number of trees, maximum tree depth, and minimum samples required for splits and leaf nodes. The fine-tuned Random Forest achieved a Test AUC-ROC of 0.741 and an Accuracy of 71.3 %, demonstrating a slight improvement in accuracy over the untuned model while maintaining a comparable AUC-ROC. Cross-validation confirmed the model's generalizability, with a mean AUC-ROC of 0.643 and a standard deviation of 0.071. Despite strong overall performance, the confusion matrix revealed persistent misclassifications, particularly for funded grants, suggesting opportunities for further refinement (Figure 3).

To further improve predictive performance, we implemented Gradient Boosting, an ensemble technique that sequentially builds decision trees by fitting each subsequent model to the residuals of the previous

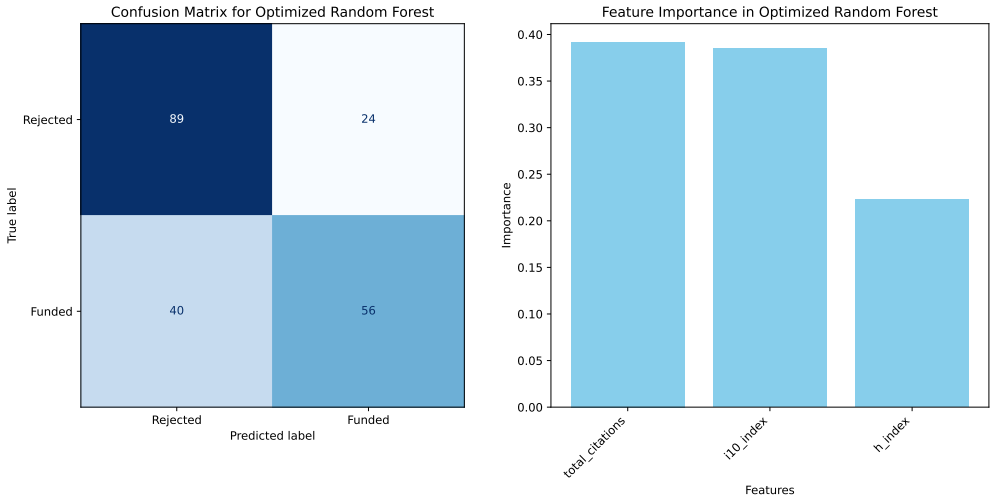


Figure 3: Confusion Matrix and Feature Importance for Fine-Tuned Random Forest.

ones. Unlike Random Forest, Gradient Boosting introduces dependencies between trees, allowing the model to minimize errors iteratively. Using a randomized search approach, we optimized key hyperparameters, including the number of boosting stages $n_{\text{estimators}}$, learning rate η, maximum tree depth, and subsampling rate. The best hyperparameters identified were $n_{\text{estimators}} = 200$, $\eta = 0.01$, *max_depth* = 10, *min_samples_split* = 5, *min_samples_leaf* = 5, and *subsample* = 1.0. The optimized Gradient Boosting model achieved a Test AUC-ROC of 0.743 and an Accuracy of 70.3 %, representing a slight improvement in accuracy over the fine-tuned Random Forest model while maintaining a comparable AUC-ROC. Feature importance analysis again highlighted total citations and i10-index as the dominant predictors, contributing most significantly to grant approval predictions, while the h-index played a smaller role. Residual analysis revealed a bimodal distribution of errors, suggesting frequent underestimation and overestimation of probabilities, particularly for borderline cases. Cross-validation results confirmed the model's generalization ability, with a mean AUC-ROC of 0.623 and a standard deviation of 0.076, though the variability across folds suggests opportunities for further refinement (Figure 4).

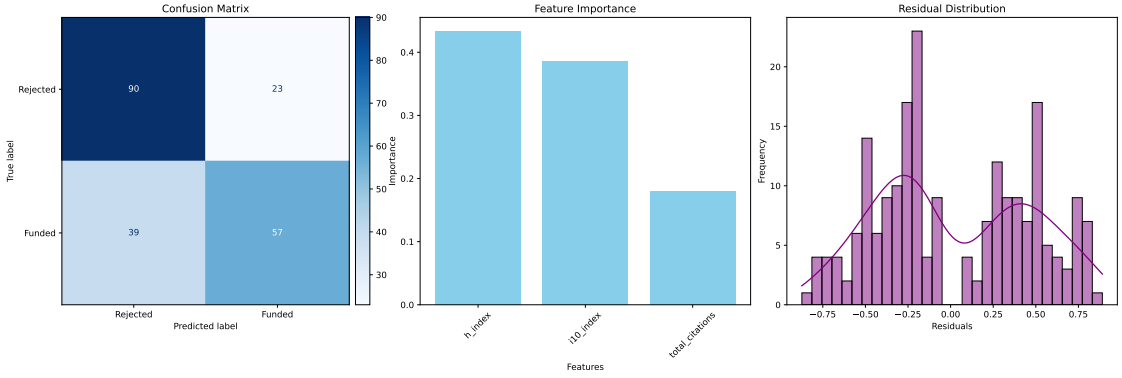


Figure 4: Evaluation of the Gradient Boosting Model: Confusion Matrix, Feature Importance, and Residual Distribution.

To compare the models tested, Table 4 summarizes their performance across Test AUC-ROC, Accuracy, and cross-validation metrics. Logistic Regression, serving as the baseline model, showed weak performance with a Test AUC-ROC of 0.570 and an Accuracy of 53.9 %. The untuned Random Forest significantly improved upon this baseline, achieving a Test AUC-ROC of 0.745 and an Accuracy of 68.4 %. Fine-tuning the Random Forest improved accuracy to 71.3 % while maintaining a Test AUC-ROC of 0.741, demonstrating its

Table 4: Comparison of Model Performance Metrics

Model	Test AUC-ROC	Test Accuracy	CV Mean AUC-ROC	CV Std. Dev.
Logistic Regression	0.570	53.9%	N/A	N/A
Random Forest (Untuned)	0.745	68.4%	0.643	0.071
Random Forest (Fine-Tuned)	0.745	71.3%	0.643	0.071
Gradient Boosting	0.743	70.3%	0.623	0.076

robustness. Gradient Boosting emerged as the best-performing model, achieving the highest Test Accuracy of 70.3 % and a Test AUC-ROC of 0.743, narrowly outperforming the fine-tuned Random Forest. Despite its strong performance, the cross-validation results for Gradient Boosting revealed moderate variability, with a mean AUC-ROC of 0.623 and a standard deviation of 0.076. Across all models, feature importance analysis consistently highlighted total citations and i10-index as the most influential predictors, underscoring their critical role in determining grant approval probabilities.

In conclusion, Gradient Boosting demonstrated the strongest predictive performance, achieving the highest accuracy while maintaining a robust AUC-ROC. However, further refinements, such as ensemble stacking methods, advanced algorithms like XGBoost or LightGBM, and additional proposal-specific features, could enhance generalization and address variability in performance.

5 Conclusion

In conclusion, our analysis reveals that diversity in proposal submission strategies does not necessarily correlate with higher funding success from the DoD. While TDA offered insights into structural patterns, it indicated that more focused approaches may be more effective. NLP-based classification using titles, keywords, and authors showed encouraging but not definitive improvements, suggesting that refining models and selecting better features could enhance performance. Finally, examining bibliometric data (e.g., h-index) hinted at the importance of researcher reputation but did not yield conclusive predictive power. Overall, the results guide future work toward more targeted strategies, refined text features, and careful consideration of academic metrics to improve funding success predictions.

References

[1] J. Roehm, *An application of tda to professional basketball [jerome roehm]*, https://www.youtube.com/watch?v=-cfp-tH-vIM, 2023.

[2] L. Bornmann and H.-D. Daniel, "What do we know about the h index?" *Journal of the American Society for Information Science and Technology*, vol. 58, no. 9, pp. 1381–1385, 2007. DOI: https://doi.org/10.1002/asi.20609. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20609.

[3] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, Nov. 2005, ISSN: 0027-8424. DOI: 10.1073/pnas.0507655102. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283832/.

[4] Google, *Google Scholar Blog: Google Scholar Citations*, Jul. 2011. [Online]. Available: https://scholar.googleblog.com/2011/07/google-scholar-citations.html (visited on 12/16/2024).