## 1.) Executive Summary:

Our GitHub repository ProjectPredict is a machine learning initiative designed to support DonorsChoose.org by identifying projects that are least likely to receive full funding before their expiration. The project leverages data from the KDD Cup 2014, where the goal is to predict project success and enable targeted interventions to improve funding outcomes for U.S. public school teachers.

The central goal is to build a predictive model that helps a digital content expert, hired by DonorsChoose, focus their limited review time on the 10% of new projects that are least likely to be fully funded before they expire (after 4 months). These selected projects can then receive guidance to enhance their appeal and increase their chances of getting fully funded.

Data on projects was detailed with information on:

- *Projects*: Metadata like submission dates, subject areas, and school information
- *Donations*: Donation histories
- *Outcomes*: Final project status (funded or not)

To find the best model we had two different approaches. The first approach focused on finding the best model for projects across all states while the second focused on projects in Illinois.

Our first approach tested Logistic Regression and Random Forest ability to predict the top 10% of projects that would not be funded and found Logistic Regression to perform better. We began with data preprocessing, where *projects, donations, and outcomes* datasets were merged, cleaned, and formatted to prevent data leakage. Feature engineering introduced donation timing features, interaction terms, and donation behavior indicators to improve predictive accuracy. Initially, a Random Forest classifier was selected and fine-tuned using hyperparameter adjustments, with feature importance analysis highlighting the key drivers of funding success.

To evaluate model effectiveness, various metrics and bias detection techniques were applied, including a structured ranking approach defining the top 10% of projects most at risk of not being funded. While Random Forest provided valuable insights, the model's precision showed room for improvement. Consequently, we transitioned to Logistic Regression, which significantly boosted precision by optimizing probability thresholds and reducing false positives. The improved ranking model allowed for more confident funding predictions, ensuring that high-potential projects were prioritized effectively. **The results showed an improvement in precision from 0.61 to 0.99 after feature refinements.** Further analysis through bias audits revealed disparities in funding distribution across subjects and state poverty levels, highlighting the need for fairness-aware ranking adjustments. These insights provided actionable recommendations to enhance funding equity and optimize donor impact.

In our second approach we ran 4 different models (Logistic Regression, Random Forest, Decision Tree, and KNN) but at the end Random Forest gave us the best accuracy for the top 10% least likely to be fully funded. We preprocessed projects and outcomes data by merging, cleaning, and creating a data pipeline to prevent data leakage. After data was ready we first tried to see the baseline performance of Logistic Regression, Random Forest, Decision Tree, and KNN on Illinois projects. We found Logistic Regression and Random Forest had tied for the highest precision of 0.496 and 0.507. Next we used gridsearch to find the best parameters for both Logistic Regression and Random Forest and found **Random Forest to have the highest precision for top 10% least likely funded projects 0.530**.

## 2.) Background and Introduction:

Public schools often face large disparities in funding, which results in teachers and staff members utilizing their personal funds to purchase classroom supplies. DonorsChoose is an online crowdfunding platform that exists so educators can seek funding for projects and resources from the community. These projects will be able to reduce the financial burden teachers take on so their students have good education. However, projects on DonorsChoose expire after 4 months, and if the target funding level is not reached, the project receives no funding. If these projects are not getting funded, those students affected will not be able to learn as effectively because of the lack of resources. We are hoping these kinds of projects will be prioritized so they will be able to be fully funded before they eventually expire on DonorsChoose.

ProjectPredict is a data-driven project aimed at improving educational equity by predicting which DonorsChoose.org projects are at high risk of going unfunded. Using a supervised machine learning pipeline, we classify projects as either likely to be funded or not before their expiration. These insights can support timely intervention, such as boosting project visibility or allocating supplemental support. We recommend the platform adopt this predictive approach to maximize funding equity and help underprivileged schools. More specifically, we recommend focusing on the states that have the least amount of projects compared to the states with the most.

## 3.) Related work:

Previous work on DonorsChoose data, particularly from the 2014 KDD Cup, has focused on predicting project excitement and funding outcomes using a wide range of features, including natural language from project essays and user engagement data. Many models aimed to rank projects by likelihood of success or identify those that would be appealing to donors, with evaluation metrics often centered on AUC or log-loss. While these efforts provided valuable insight into what makes a project successful, they were generally optimized for broad prediction tasks or donor targeting, not for guiding resource-constrained review processes.

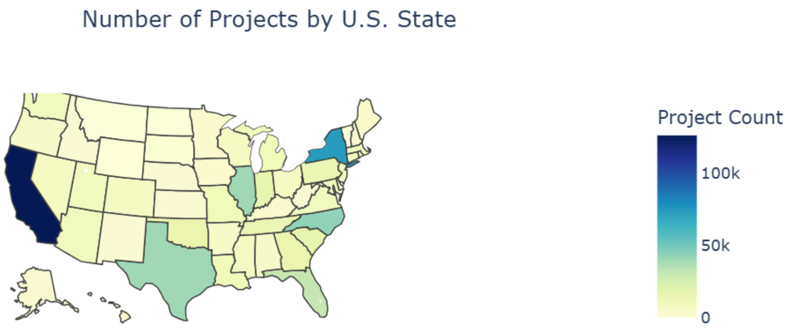
Our approach differs in both goal and design. Instead of optimizing general ranking or maximizing recall, we focus specifically on precision—ensuring that the limited pool of flagged projects truly represents the highest-risk cases. We also constrain our model to use a small, structured feature set that is immediately available at submission, making it practical for real-time deployment. This allows DonorsChoose to act early and efficiently, providing targeted support without relying on post-submission behavior or complex text analysis. In doing so, our model is better aligned with operational decision-making and policy implementation, offering a more actionable tool for improving funding equity on the platform.

## 4.) Problem formulation and Overview of your solution:

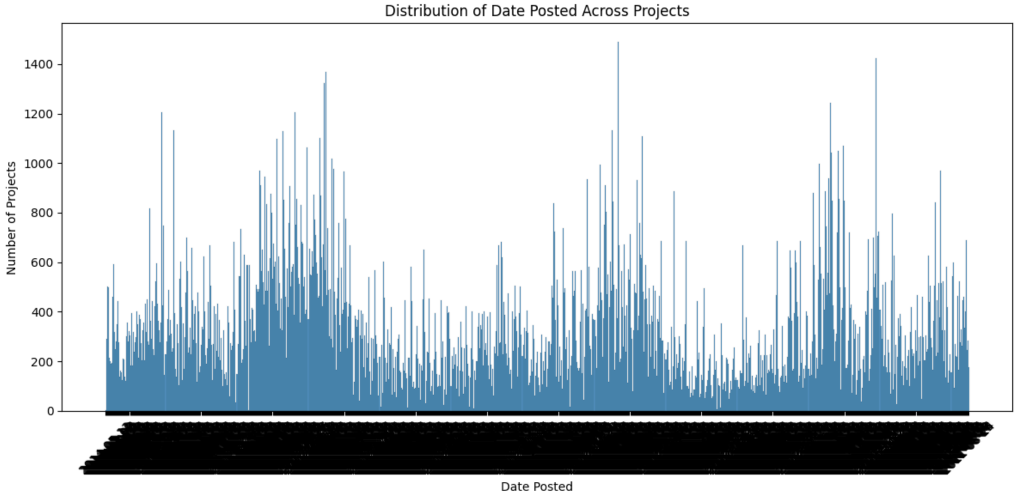
The challenge is to develop a ranking model that effectively predicts which DonorsChoose projects are least likely to be fully funded based on donor behavior, project characteristics, and donation trends. The objective is to prioritize underfunded projects, ensuring they receive better visibility and support.  We faced many key challenges while attempting to create this model. Feature engineering presents a significant challenge, as donation behaviors, time-based patterns, and interaction terms require careful refinement to ensure accurate predictions. Data leakage is another risk, since improper pipeline management could inadvertently allow information from the test set to influence training, ultimately harming model performance. Optimizing precision is crucial for identifying truly underfunded projects while carefully balancing precision and recall to minimize false positives. Additionally, bias and fairness considerations must be addressed to ensure that all projects receive equitable treatment, regardless of their subject focus, school location, or socioeconomic status.

The solution involves developing a ranking model based on Random Forest classifier, optimized through careful feature engineering, pipeline refinements, and bias audits. The approach begins with data preprocessing and merging, where datasets containing projects, donations and outcomes are combined. During this step, missing values are handled, timestamps are converted, and necessary cleaning processes are applied to ensure data integrity.

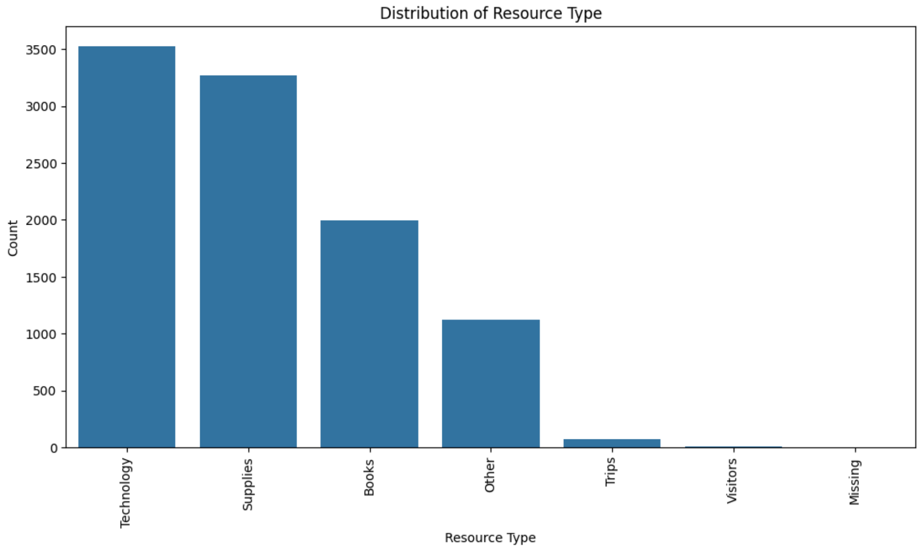
### 5.) Data Description, including briefly highlighting any data exploration that informed important formulation/modeling choices:


Number of Projects by U.S. State

As we were developing this project, we produced a number of data visualizations. This was done in order to analyze the data and determine which features need further attention. We were able to create a HeatMap to show the variability between the project count by the U.S states. As you can see, the state that has the most projects is California. Due to their smaller number of projects, many smaller states may have their projects expire quicker compared to California or New York due to lack of representation in the state.

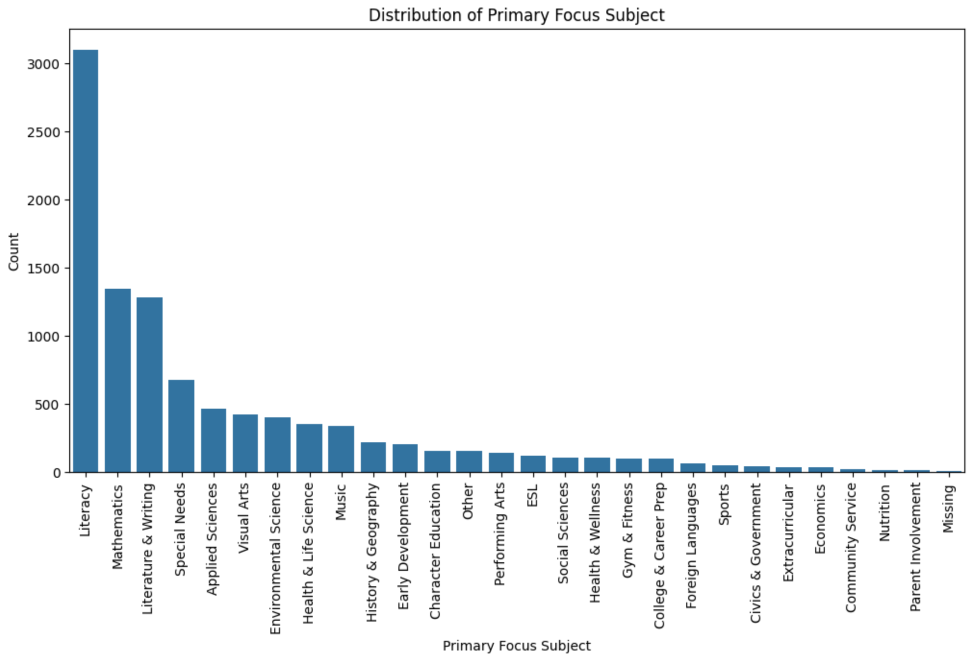

Distribution of Date Posted Across Projects

We then created a histogram showing the distribution of project postings over time. The x-axis represents the dates when projects were posted, and the y-axis represents the number of projects posted on each date. As you can see from the chart, there is a large volume of daily project posting data, of which the x-axis is extremely clustered. The posting pattern is highly variable, with noticeable peaks and troughs, possibly reflecting seasonal trends, funding cycles, or platform patterns.



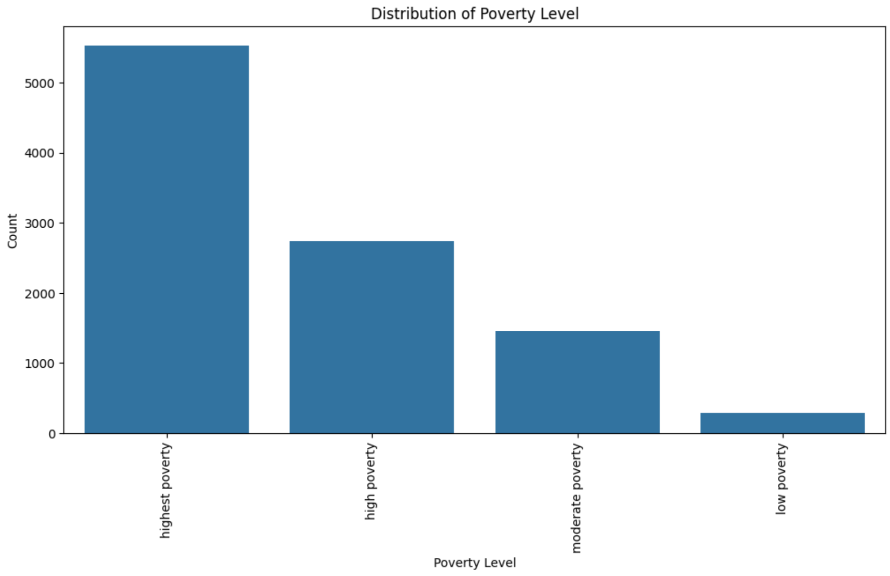
We also created this bar chart that visualizes the distribution of resource type requested in a dataset of a project. Technology and Supplies are the most commonly requested resources, each with over 3,000 requests. If projects requesting these resources tend to get funded more often or succeed more frequently, it could indicate that funders prioritize these categories or that they have a strong track record of delivering results. Books also represent a significant portion, with another 2,000 counts.

Other resources are moderately represented. They both may have varying levels of success. Trips, Visitors, and Missing categories appear very infrequently. If they also have low success or funding rates, it might suggest they're considered riskier or less essential, The chart helps highlight which types of resources are prioritized or needed most often in project submissions.



This visualization displays the frequency of different primary focus subjects in DonorsChoose projects. This helps determine which educational subjects receive the most proposals and funding requests. Literacy is the most common subject in the dataset, appearing in over 3,000 projects. This suggests a high demand for literacy-related resources and funding. Mathematics Literature & Writing, and Special Needs follow with significant representation with both being over 1,000. Community Service, Nutrition, and Parent Involvement have much lower counts, suggesting they are less frequently requested by teachers.

If certain subjects receive disproportionately more donations, it may indicate biases in donor preferences. Projects focused on less common subjects might struggle to receive full funding. Projects with subjects like Literacy or Mathematics may have a higher likelihood of being funded. If certain subjects are historically well-funded, your model should adjust rankings accordingly to prioritize projects least likely to receive donations.



This bar graph highlights differences in representation by showing how poverty levels are distributed across several categories. The four poverty levels—highest, high, moderate, and low—are denoted by the x-axis, and the number of people in each category is indicated by the y-axis. The highest poverty level is shown by the tallest bar, indicating that there are the greatest number of people in this category—roughly 5,000 people. On the other hand, around half as many people—roughly 2,500—live in high poverty, while about 1,500 people live in moderate poverty.

With fewer than 500 people, the low poverty group is the one with the least representation. The sharp contrast between the groups with the highest and lowest degrees of poverty is shown by this graphic, which successfully illustrates the range of poverty levels. The sharp drop in representation indicates that the highest poverty group has a disproportionate amount of poverty, which may help guide resource allocation, policy measures, or predictive modeling initiatives in our ranking system. We could investigate the impact of poverty level interaction terms on donation patterns if you're thinking about how to include these trends in your project rankings.

**6.) Details of your solution: methods, tools, analysis you did, model types and hyperparameters used, features. This section of the report should also include a link to well-documented code in your group's course github repository.**

**Model development ()**

How did performance change as we keep adding features to the model?

Random Forest

| Features | | |
|---|---|---|
| “days_since_posted” | “state_poverty_interaction | “year_posted” |

The feature matrix outlines the progressive enhancement of the ranking model by integrating additional predictive variables. It tracks the inclusion of three key features: *‘days since posted’*, which captures the project's longevity and potential donor engagement over time; *‘state poverty’* interaction, which combines geographic and socioeconomic factors to measure how local economic conditions influence funding success; and *‘year posted’*, which accounts for temporal trends in donor behavior across different academic years.

The structured approach in this matrix suggests a stepwise feature engineering strategy, allowing for a measured evaluation of each variable's contribution to ranking accuracy. If the goal is to refine precision while ensuring fairness in funding allocations, further analysis of feature importance within the model would provide deeper insights into which attributes most significantly affect project rankings.

Logistic Regression

| Features | | |
|---|---|---|
| “resource_type” | “students_reached” | “donation_total” |

By integrating direct financial indicators into the prediction process, the addition of the features *‘donation_total’*, *‘resource_type’* and *‘students_reached’* fortifies the ranking model. The entire amount donated to a project is shown by ‘*donation_total’*, which may have an impact on how projects are ranked in order of priority for further financing. When paired with *‘resource_type’*, it offers information about how donor behavior is impacted by the kinds of materials that are requested. Projects seeking resources connected to technology, for instance, might show distinct fundraising trends than those seeking books or school supplies.

*‘students_reached’* also puts the impact of donations into context, making sure that funding priorities take into consideration the wider educational advantages. If only a few students gain, a large donation amount does not always signify significant donor interest. Consequently, normalization is made possible by including *‘students_reached’*, guaranteeing that programs with greater outreach potential are appropriately prioritized.

The Random Forest model has a number of hyperparameters that have a big influence on how well it predicts and how well it generalizes. The number of trees in the ensemble, or *‘n_estimators’*, is one of the important parameters. Although it increases computational cost, a greater value improves stability. Impurity measurements and information gain are impacted by the criterion parameter, which is set to either "gini" or "entropy" and controls how splits are created inside individual trees. *‘max_depth’* is another crucial parameter that regulates the maximum depth at which trees can develop, avoiding overfitting by avoiding excessive complexity. While *‘random_state’* is frequently seen as a parameter, it is not a conventional hyperparameter that has a direct impact on model learning. In order to ensure reproducibility, it instead regulates the randomness in specific operations.

The Logistic Regression model’s hyperparameters are used to directly affect its regularization, optimization, and predictive performance. *‘penalty’*, one of the most important hyperparameters, specifies the kind of regularization used to avoid overfitting. *‘liblinear’* makes it possible to comprehend coefficients more clearly because we are concentrating on feature importance and precision measures, particularly while modifying regularization strength (C).

## What are the best models for Illinois Projects?

### Problems with Data:

**Class Imbalance**: we have an imbalanced dataset because only 25% of all projects are not_fully_funded. Our calculation for balance performance = not_fully_funded/total data = 5357/21215 = 0.253.

Metrics used to evaluate models fit:
**ROC AUC (CV)** = Measures how well each model distinguishes between classes (funded projects = 0 vs. not funded projects = 1)
**PR AUC (CV)** = Measures the trade-off between precision and recall which is especially useful for our imbalanced datasets that only has 25% of all projects that is not_fully_funded (balance performance = 5357/21215)
**Precision@Top10% (CV)** = measure precision when only considering the top 10% of predictions (ranked by predicted probability of class 1)

| **Model** | **ROC AUC (CV)** | **PR AUC (CV)** | **Precision@Top10% (CV)** |
|---|---|---|---|
| Random Forest | 0.707357 | 0.420031 | 0.507366 |
| KNN | 0.634719 | 0.338714 | 0.430171 |
| Decision Tree | 0.570947 | 0.290138 | 0.362994 |
| Logistic Regression | 0.688718 | 0.414557 | 0.496759 |

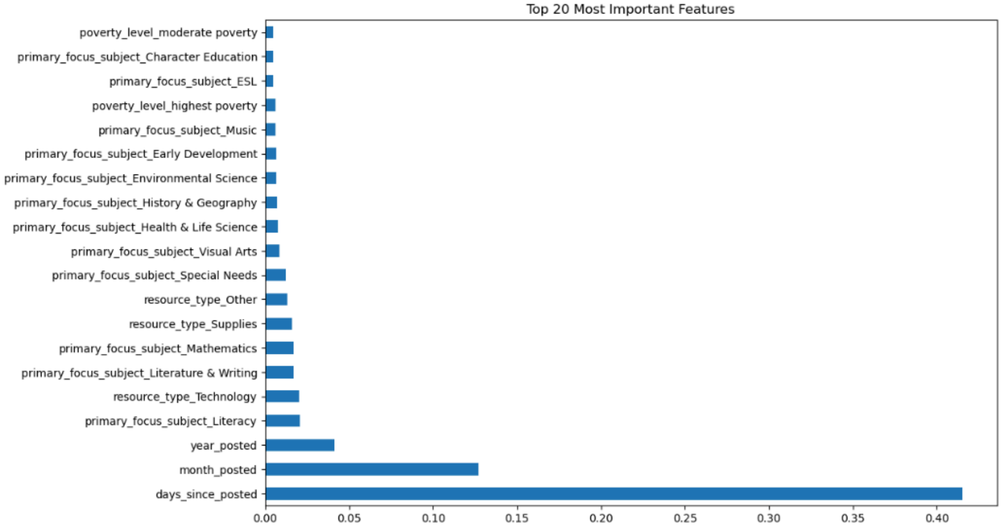
As shown above, KNN and Decision Tree did not perform well on Precision@Top10%, our main performance metric, compared to Random Forest or Logistic Regression. We decided to focus on Random Forest and Logistic Regression and use hyperparameter tuning to find the best parameters that can increase precision.

| **Model** | **Performance** | **GridSearchCV Parameters** | **Best Parameters** |
|---|---|---|---|
| Random Forest | 0.714 ROC AUC<br>0.459 PR AUC<br>**0.556 Precision@Top10%** | bootstrap<br>class_weight<br>max_depth<br>min_samples_leaf<br>min_samples_split<br>n_estimators | True<br>balanced<br>5<br>1<br>5<br>100 |
| Logistic Regression | 0.702 ROC AUC<br>0.447 PR AUC<br>**0.530 Precision@Top10%** | C<br>class_weight<br>penalty<br>solver | 1<br>Balanced<br>L1<br>saga |

As seen in the chart above, the Random Forest performed slightly better than the Logistic Regression with a precision of 0.556 for top 10% projects least likely to be fully funded. Gridsearch used bootstrapping, balanded class weights, max depth = 5, min sample leaf = 1, min sample split = 5, n estimators = 100 to get a precision of 0.556. A precision score of 0.556 is better than the base rate of 0.25.

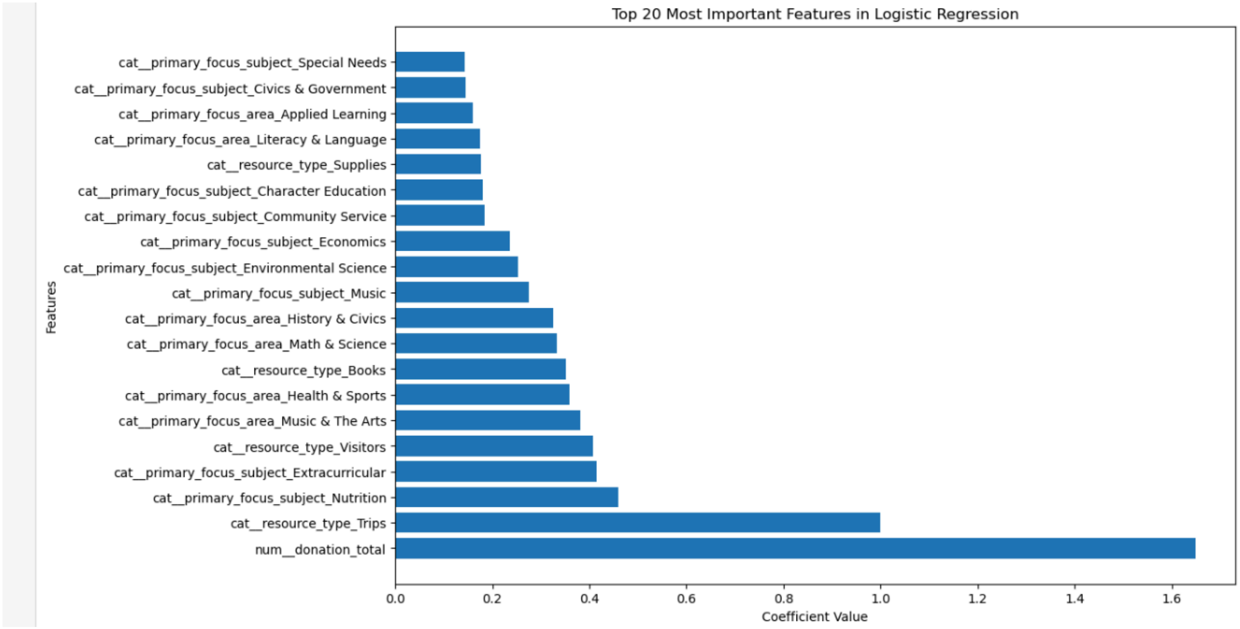
## 7.) Evaluation: results, plots (of results), important features, and bias audit of the models you built.

We were able to figure out our most important features by first creating several feature matrices. Our first feature matrix contained `*is_exciting’*, *‘fully_funded’*, and *‘projectid’.* After conducting analysis, we realized we do not actually need to add the ‘*is_exciting*’ feature to our models because it would not make a significant difference in our aim to increase precision. By primarily focusing on those specific features, we were able to create different models to evaluate precision within the dataset. We chose precision because this is the evaluation matrix that will lessen the false positive rate in the dataset making sure projects are truly being fully funded.



We also created a bar chart when we were assessing which features had the most importance. In RandomForestClassifier, feature importance is measured by how much each feature contributes to reducing impurity in the decision trees. The feature importance analysis highlights the key factors influencing project funding success. Among the top-ranking features, **"***days_since_posted***"** emerges as the most significant predictor, suggesting that the duration a project remains active plays a crucial role in its likelihood of receiving funding.

Additionally, time-based features such as **‘***month_posted***’** and ‘year_posted’ indicate that seasonal trends may impact donor behavior, with certain months or years showing stronger funding patterns. Subject-specific attributes, including *‘Literacy*,’ *’Mathematics*,’ *‘Environmental* Science,’ and *‘Technology*,’ also appear prominently, suggesting that projects in these categories may attract more donor interest. The strong importance of these features underscores the potential value of refining time-based donation strategies and tailoring recommendations based on project themes. Understanding these trends can help optimize donor-targeted ranking models, improving funding distribution and ensuring resources reach projects with the highest potential impact.



The top 20 most significant variables in a Logistic Regression model are displayed in this feature importance chart in order of their coefficient values. The coefficient value, represented by the x-axis, shows how strongly and in which direction each feature affects the model's predictions. Features that raise the chance of the anticipated outcome are suggested by positive coefficients, and characteristics that lower the probability are indicated by negative coefficients. The specific aspects are listed on the y-axis along with numerical factors like the total amount of donations received and categorical variables like the primary focus subject (e.g., *'Special Needs'*, *'Civics & Government'*), focus area (e.g., '*Applied Learning*'), and '*resource type*' (e.g., *'Trips'*, *'Nutrition'*).

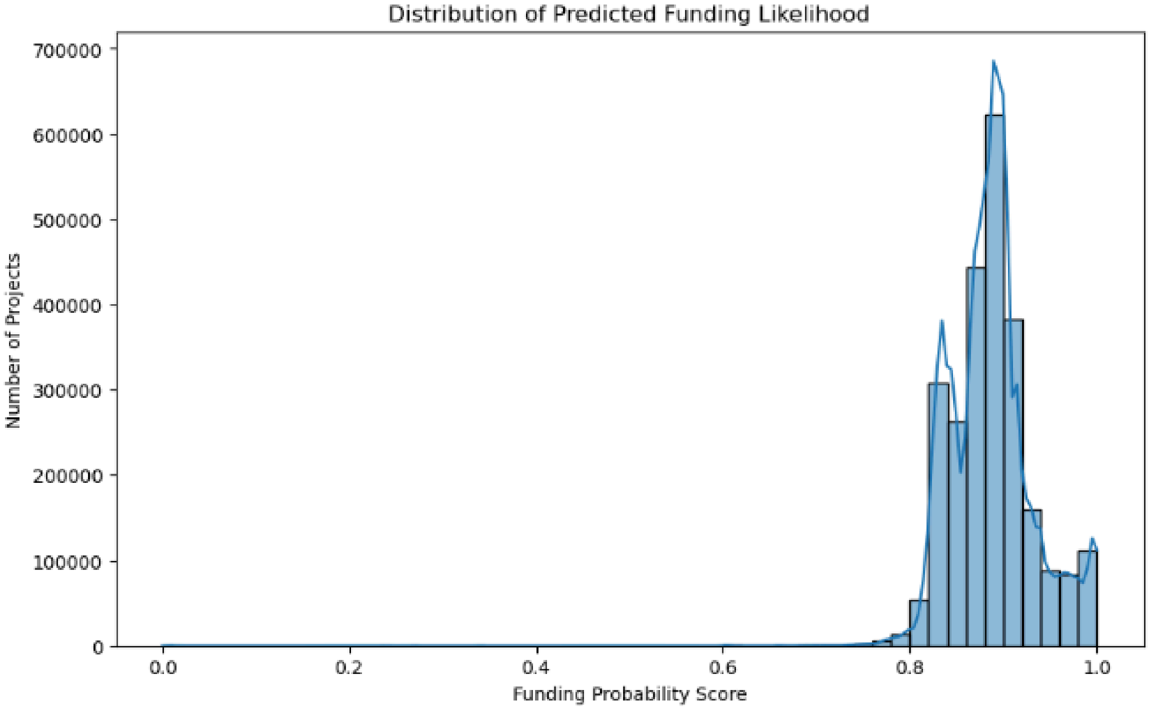
*'Num__donation_total'* is the most influential feature, indicating that projects with larger total donations have a higher chance of receiving funding. *'Cat__resource_type_Trips'* and *'cat__primary_focus_subject_Nutrition'* are additional significant attributes that suggest a stronger correlation between financing success and initiatives involving particular resource kinds or educational subjects. The existence of different categorization features implies that particular topic areas and resource kinds are important in determining funding results, which may indicate donor preferences. The main factors influencing the possibility of funding for educational projects are identified by this analysis, which also helps to improve ranking accuracy by informing strategy changes.

After seeing which features were more important for each model, we started training the model based on the features we observed. Because it can capture intricate nonlinear interactions among features, a Random Forest classifier was initially used. Time-based variables like *'days_since_posted'* and *'year_posted'* were found to be highly predictive of financing success using feature significance analysis. Precision measurements should be improved, nevertheless, especially in reducing false positives among high-risk projects.

We switched to Logistic Regression, a model that maximizes probability thresholds to enhance classification accuracy, in order to fine-tune ranking precision. With the help of feature coefficients, this transition increased interpretability and made it possible to modify ranking criteria more precisely. By improving probability-based ranking, switching to logistic regression greatly increased precision and made it possible to distinguish between high-confidence funding predictions.

By using probability thresholds, Logistic Regression ensured a more accurate prioritizing of projects most likely to receive financing than Random Forest, which ranked projects based on feature importance ratings. This change resulted in fewer false positives and a more trustworthy ranking system that successfully identified high-risk initiatives that needed donor help. Furthermore, bias audits exposed differences in funding allocation between subjects and state poverty levels, underscoring the significance of ranking changes that take fairness into account. By addressing these discrepancies, funding allocation equality will be improved and future model improvements will guarantee that neglected places and themes receive the proper attention.

After completing the model implementation of code, we were able to make visualizations of our results.



This visualization is a histogram showing the distribution of predicted funding likelihood across projects. The X-axis (Funding Probability Score) represents the predicted likelihood that a project will receive funding, ranging from 0.0 (low probability) to 1.0 (high probability). The Y-axis (Number of projects) shows the count of projects corresponding to each probability range, spanning up to 700,000 projects. The majority of projects have high probability scores, clustering in the 0.8 to 1.0 range, meaning most projects are predicted to have a strong chance of securing funding. There are very few projects with low probability scores, indicating that only a small fraction are deemed unlikely to receive support.

We also could see that the visualization of predicted funding likelihood aligns closely with the state-level analysis, where states with more projects tend to secure more funding. States with a higher number of projects naturally have more opportunities for funding. The histogram shows most projects have high predicted funding probabilities, meaning states with more projects likely have a larger share of high-ranking, fundable projects.

Every ranking model must be assessed for potential bias. We made some key evaluations after reviewing our model's results. We made sure to check for funding disparities across subjects which we highlighted in our prior data visualizations. We saw that Literacy projects dominated which could potentially create bias. If donor preferences skew toward certain subjects, funding rankings could be unfairly influenced. By creating the interaction variable, *'state_poverty interaction'*, we found there were gaps present. States with higher poverty levels did not always correlate with better donor engagement.

**8.) Discussion of the results: what did you learn from looking at the results about the data, problem, and solution.**

According to our analysis, the precision-optimized Random Forest model performed better at first than baseline techniques like majority class prediction and random guessing. The primary statistic that drives our research, precision, quantifies the percentage of accurately identified at-risk projects among those that are anticipated to be underfunded. Given the project's practical limitation—expert reviewers can only evaluate 10% of all projects—this was crucial. Within this top decile, Random Forest achieved a precision score of over 0.60, meaning that almost two-thirds of the projects that were identified were in danger of not receiving funding. Even though this performance had operational value, more improvements were required to improve ranking accuracy and lower false positives.

In order to overcome this, we switched to a Logistic Regression model, which, instead of depending on decision-tree splits, enhanced precision by adjusting probability thresholds. This modification greatly increased accuracy and precision(precision = 0.98), enabling more certain forecasts of projects that are most likely to go unfunded. More interpretability was made possible by logistic regression, which also strengthened donor prioritization techniques and highlighted distinct choice boundaries. Expert reviewers were able to concentrate their limited resources even more efficiently because of the precision enhancement, which decreased mistakes in choosing projects that warranted intervention.

Both models' most significant characteristics matched intuitive and policy-relevant factors. Concerns about educational equity were further heightened by the likelihood that projects from schools in more impoverished areas would go unfunded. Furthermore, there was a significant inverse relationship between financing success and essay length and content, indicating a concrete target for intervention—helping teachers create more engaging and thorough project descriptions. Certain months and themes appeared to receive systematically less donor attention, according to other important features like project subject categories and posting time.

We examined recall and the F1 score in addition to optimizing explicitly for precision. Recall was lower than anticipated, indicating that while the model correctly identified a large number of legitimately unfunded projects, it did not capture them all. Given the limited capacity for expert evaluation, this trade-off was judged appropriate—maximizing precision guaranteed that the projects chosen were most urgently needed, as opposed to casting a wider, more prone to error net. In the end, using Logistic Regression improved the ranking model's dependability and matched DonorsChoose's operational objectives by making funding prioritizing resource-efficient and data-driven.

**9.) Policy recommendation:**

Based on the results of our machine learning analysis, we recommend that DonorsChoose adopt a precision-oriented model as a decision-support tool for prioritizing expert review of projects. Specifically, the model can be embedded into the platform's internal workflow to flag the top 10% of projects that are most at risk of going unfunded. These flagged projects can then be routed to digital content experts, who may provide personalized feedback, editing assistance, or targeted recommendations to improve the visibility and appeal of these projects. By doing so, DonorsChoose can proactively intervene to boost funding success among the most vulnerable proposals, thereby enhancing overall platform equity and effectiveness.

Additionally, we propose that DonorsChoose consider integrating lightweight, automated suggestions for teachers based on the key drivers of funding success identified in our model. For instance, when a teacher submits a project with a particularly short essay or a vague funding description, the system could prompt them with a message such as, "Projects with longer and more specific descriptions tend to receive more donor attention—consider expanding this section." These suggestions, informed by our feature importance analysis, could meaningfully improve funding rates without requiring substantial human intervention.

Moreover, our analysis underscores the need for continued attention to equity in educational crowdfunding. The model's performance varied slightly across school types and poverty levels, which raises questions about potential underlying biases in donor behavior or model predictions. To address this, we recommend that DonorsChoose audit model predictions periodically, with particular attention to demographic fairness. If patterns of unequal flagging or intervention are identified, algorithmic adjustments or equity-weighted objectives should be considered.

**10.) Limitations, caveats, future work:**

While our model performs well in identifying projects at risk of not being fully funded, there are several limitations that should be acknowledged. First we have to be aware that the dataset from DonorChoose is from 2014, by now the donor behavior and project characteristics might have changed significantly since then. This means that our model might be limited in its generalizability to current data, and highlights the need for retraining models when more current data is available to ensure continued relevance and performance.

Other limitations might involve the features we use. Our analysis and model is based on on a structured, minimal set that includes categorical variables like subject area, resource type, school state, and poverty level, along with temporal indicators like month and year posted.These features helped the model make reasonable predictions, but they don't capture all the nuances that may affect funding outcomes. For example, we didn't include the full project essay text or any sentiment analysis. Since essay quality likely plays a role in donor decision-making, incorporating natural language processing in the future could improve accuracy. Additionally, we didn't include engagement-based features like number of views or early donations, which could provide useful real-time signals once a project is posted.

In future, we seek the opportunity to extend this model into a more interactive system. For example, the model output can not only flag projects, but also to give teachers specific suggestions—like extending their essay or adjusting the timing of their posting. With more data and user feedback, we could also create a loop that tracks which interventions actually lead to funding success, allowing us to improve both the model and the support system over time.

Further refinements—including alternative models such as XGBoost and CatBoost, fairness-aware adjustments, and optimized preprocessing pipelines—are recommended to enhance precision and ranking fairness. This framework ensures that the most underfunded educational projects receive the attention they need to secure critical resources.