Optimized Crowdfunding

Final Report

Group I

5/2/2016

Name	Andrew ID
Dhruv Bhogle	dbhogle
Emilio Esposito	eesposit
Mark Minisce	mminisce

- Objective
- Introduction
- Business Motivation
- Data Understanding and Scope
  - Lending Club Dataset
  - Kickstarter Dataset
- Technical Approach
  - Loan Risk Assessment Model
  - Fundraising Model
  - Savings Estimation Discussion and Illustration
- Results and Discussion
- Risks and Risk Mitigation Strategies
  - Text Features Increase Risk of Overfitting
  - Limited Common Data Across Various Crowdfunding Platforms
  - Optimization Parameters Outside the Borrower’s Control
- Future Work
- References

Objective

This project will attempt to solve a data driven problem by integrating data gathered from P2P lending platforms, such as Lending Club (https://www.lendingclub.com/), and crowdfunding donation platforms, such as Kickstarter (https://www.kickstarter.com/). We will aim to achieve the following business objectives:

1. Reduce financial burden on borrowers by investigating multiple funding sources.
2. Increase the probability a borrower is approved for a loan.
3. Reduce lending and securitization risk for loan originators.
4. Grow the third party lending business as a means for investment.

Introduction

Many individuals in our society face financial hurdles that can inhibit them from achieving a goal. In these instances, people often appeal to the crowd to seek financial assistance through loans or donations. As shown in Fig 1, both these modes of financial assistance have experienced a comparable increase in growth rate since 2010. There are several existing online crowdfunding platforms that allow individuals to lend or donate money; however, there is not a mechanism that allows a borrower to optimize their fundraising using a mix of both donations and loans from these platforms. Our solution addresses this issue by presenting a prospective borrower with a mix of options to maximize the likelihood of securing donations while minimizing risk on borrowed funds.These objectives reduce the financial burden for an individual seeking viability for an idea and encourages innovation.

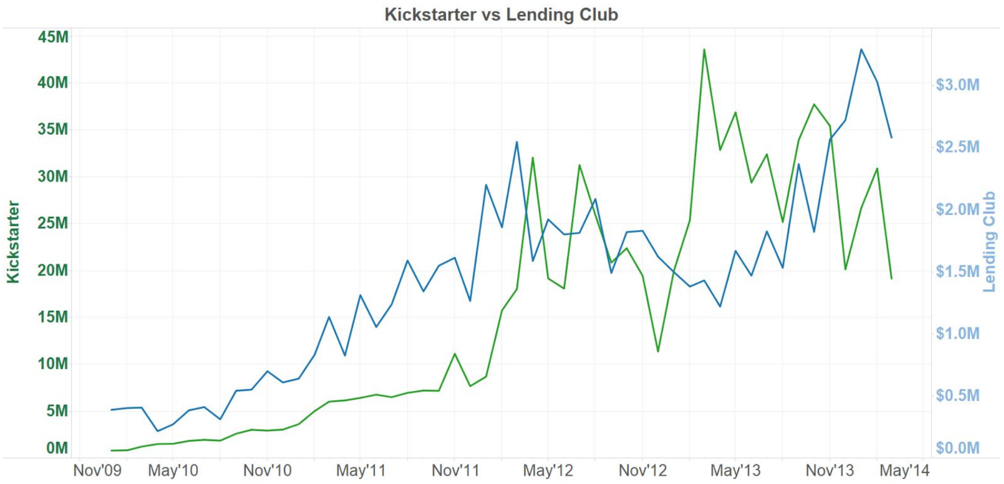


Fig 1. Financial trends in Kickstarter and Lending Club from 2010 – 2014.

Business Motivation

There are three primary groups of beneficiaries of our system. Loan consumers, the individuals who are seeking funds from crowdfunding sites, will be able to estimate the amount of funds they can receive from a variety of funding sources. By investigating donation funding sources, our system will reduce the total loan amount and interest paid by individuals. Our system will reduce individual loan default risk originators such as Lending Club bear by reducing the amount sought after by consumers. Finally, investors in securitized loan packages offered by loan originators are provided with more reliable income streams through the investment in more stable investment vehicles.

Our solution provides valuable information that loan consumers can use before they seek funds. Specifically, we are targeting individuals that have the option of raising money through either a donation or lending platform (e.g. small business funding). There is a cost and a risk associated with each application to raise funds from either of these sources. Each time a person applies for a loan, it creates a "hard inquiry" to their credit history which can hurt their chances of receiving future loans by increasing their perceived risk on the credit market, and thus their interest rate. Similarly, if an individual asks for too much money in a donation campaign attempt, potential donors will be more skeptical of the individual's future campaigns. Knowing the likely outcomes ahead of time would enable them to optimally allocate the amounts they seek from each platform on their first attempt. In the future, our work could help aspiring entrepreneurs or individuals with sensitive credit explore financing options with an online platform.

Loan originators can create investment products by selling securitized loans to investors. The securitized investments or "funds" consist of many individual loans of various risk ratings ranging from A-G (A being the most favorable and indicating the least risk). The average risk level of individual loans within each fund is dependent on the goals fund and the target investor market. By presenting donation funding first, our system will reduce the total amount of dollars lent to all consumers, reducing the aggregate risk to originators. Furthermore, originators will more easily be able to control the risks of their investment offerings by having more high quality loans available for securitization. Finally, our system will reduce the risk undertaken by investors in securitized loan packages by improving the average risk grade of individual loans in each investment. By increasing investment quality, our system will further legitimize the third party lending market in the eyes of investors.

Data Understanding and Scope

We have used two different datasets throughout the course of this project.[1] The project scope could be expanded if more high quality crowdfunding datasets were secured.

[1] Data obtained from https://crowdfunding.haas.berkeley.edu/

Lending Club Dataset

The Lending Club dataset contains records of every loan originated from 2007-2015, over 500,000 loans. Of these loans, we will filter on categories that are comparable to Kickstarter type campaigns (e.g. business loans), still leaving at least 10,000 loans (likely much more depending on final categories selected). There are over 50 attributes in the data describing the borrower (including credit rating data) and the loan (including final risk rating).

The dependent variable we are interested in predicted is the risk rating letter grade (values “A” through “G”). “A” is the most favorable rating with lowest associated interest rate, while “G” is the worst. The incidence distribution is shown in the left of Fig 2.

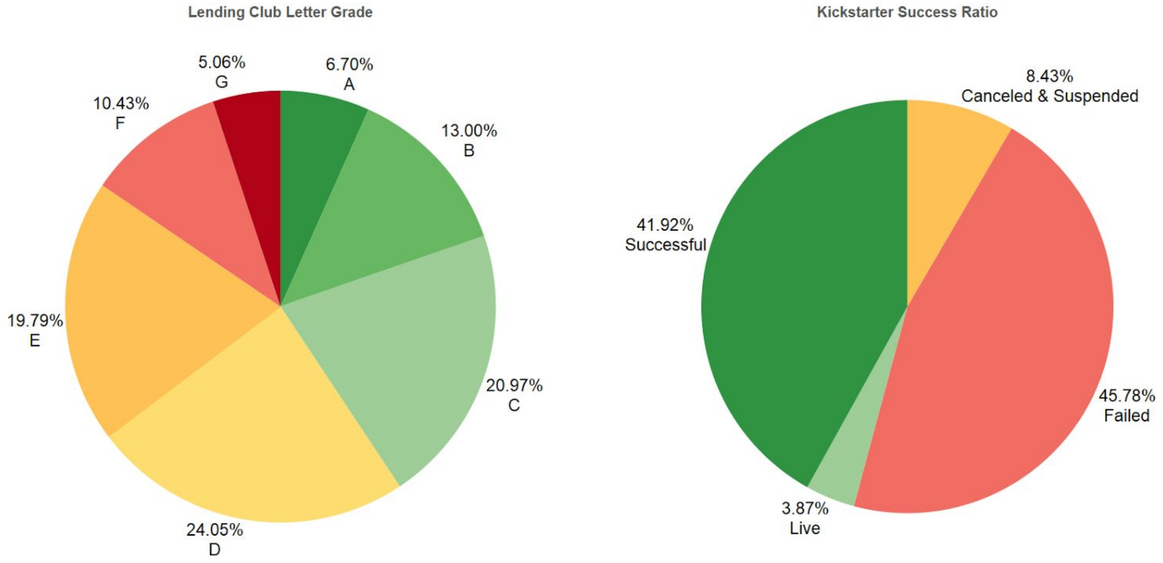


Fig 2. Incidence distribution of Risk Letter Grade (Lending Club) and Campaign Status (Kickstarter)

Kickstarter Dataset

The kickstarter dataset contains over 100,000 fundraising campaigns but is limited to 7 features: the category of the campaign, the target amount of the campaign, a free-form description of the campaign written by the requester, the home state/region of the person seeking the funds, the amount actually raised, the campaign duration and the whether or not the campaign was classified as a success. The incidence of the campaign state is shown on the right hand side of Fig 2.

In order to avoid using future information throughout our analysis, we omitted the actual amount raised by the campaign because it was highly correlated with whether or not the campaign was successful.

Technical Approach

A two component model ensemble was built and integrated to achieve our business objectives. Our first model was used to classify the risk rating of a borrower on Lending Club. Given the amount he is seeking and demographic information, the classifier returns a risk rating on the scale of A-G where A indicates the lowest lending risk. Second, we built a binary classifier on Kickstarter to predict the probability an individual will get fully funded with a confidence level of his choosing. This prerequisite is important because fundraising sites such as KickStarter require that your campaign be fully funded before distributing any money raised. By allowing the user to specify his confidence level, our model can be made more or less conservative, expanding the applicability of its use. We have used the results from the component models to estimate the savings to a prospective borrower by seeking donations instead of completely relying on a loan. The following sections detail the approaches used to build our 2 component model ensemble and illustrate our savings estimation algorithm. The performance of each component model is discussed within the “Results and Discussion” section.

Loan Risk Assessment Model

For the Lending Club risk rating multi-classifier (A through G values), several modeling approaches were attempted, including Bayes’ nets, random forests, linear regression, and an ensemble of 7 binary logistic regression models. Since the dependent variable of interest was ordinal, we first simply tried mapping the values A-G into numbers 1-7 and use a regression to predict the risk rating. However, this yielded suboptimal results when benchmarked against the performance of our 7 binary classifiers.

For feature selection, we used a combination of domain knowledge and stepwise logistic regression for each of the 7 binary classifiers. We will discuss the voting scheme of the 7 models in the Results and Discussion section. Some of the common features among the 7 models were: loan amount requested, delinquency history, debt-to-income ratio, FICO score, and home ownership status (Rent vs Own). All models were built and tested using 10-fold cross-validation.

Fundraising Model

Several approaches were used to arrive at a final probabilistic binary classifier used to predict whether or not a prospective borrower will receive 100% of the funds he requests on a crowdfunding donation platform. We attempted to fit logistic regression, naive bayes and adaboost to our data and produced two of sets of input features on which to train our models. The first set of input features consisted of the category for which funding was received, fundraising goal amount, home state/region of the donation requestor and fundraising campaign duration[2]. The second set of features contained all the aforementioned features and a set of

[2] The requester can specify the ending date of his fundraising campaign.

bi-grams generated from the project description free form field. For the text analysis, we employed the TF/IDF algorithm to identify the most commonly occurring bi-grams for campaigns that were classified as successful or failed. We then limited this to the 500 most popular bi-grams which were used as additional binary features in our model.

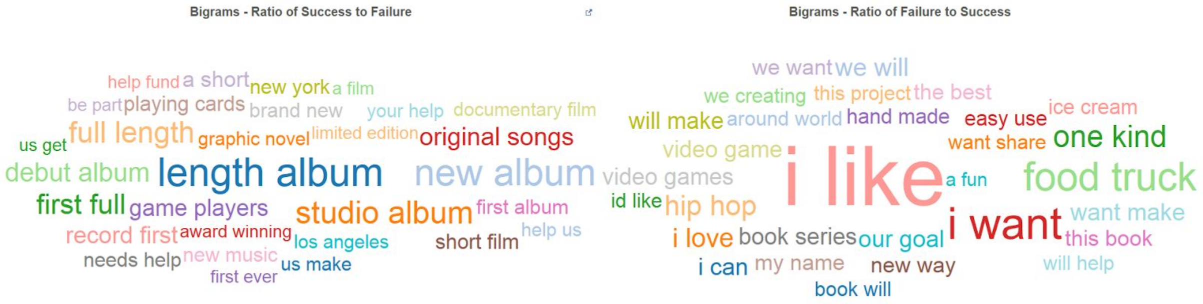


Fig 3. Popular Bi-grams in Successful (left) and Failed (right) Kickstarter Campaigns.

Descriptively, these bi-grams shown in Fig 3 tell an interesting story, successful campaigns (shown on left) tended to revolve around the music industry, one of Kickstarter's most popular domains. However, the sparseness of the dataset limited the effect of these features on our final model. To that end, the first set of input features performed better with every model while the bi-gram feature set performed worse than a default model that simply predicts a successful campaign according to the prior incidence of occurrence. Finally, Adaboost was our top performing model and therefore used in our final solution.

Savings Estimation Discussion and Illustration

In addition to the component models, we built an iterative algorithm that will allow a user to specify his level of confidence in receiving running a successful fundraising campaign. The system will continually test lower target funding amounts until the user's desired confidence level is reached. For example, if an individual seeking $50,000 wishes to be 80% confident he will receive his full ask amount, the system will continually predict the probability of full funding with subsequently lower target amounts until the desired level of confidence is reached.

Upon meeting the threshold level of confidence, the individual's risk rating will be re-estimated assuming he will be able to run a successful fundraising campaign. All information about the individual remains the same except the donation funding is subtracted from his initial loan amount. Theoretically, if the individual approaches a lender for less money, his perceived risk will decrease, improving his risk rating and decreasing his cost of borrowing. An individual seeking $50,000 for a small business loan might initially be given a B[3] risk rating. If he is able to receive $25,000 during a fundraising campaign, he may be perceived as less risky, warranting an A risk rating. The total interest savings can be calculated using the difference in interest rate ranges associated with the assigned risk ratings before and after a fundraising campaign is completed. Throughout our analysis, we have computed the interest rate savings using the third

[3] Risk ratings range from A-G. A indicates the lowest level of risk

and first quartiles of interest rates associated with a particular grade to give the user the most probable amount of savings that will result from their improved risk rating. The range of interest rates for risk ratings A and B for this example are shown in Table 1.

Table 1. First and Third Quartile Interest Rate for A and B Risk Ratings.

Risk Rating	1st Quartile Interest Rate	3rd Quartile Interest Rate
A	6.62%	7.90%
B	10.37%	12.12%

If the individual discussed above seeks a 3 year loan[4] and his risk rating improves from B to A as a result of reducing his target amount from $50,000 to $25,000 he will be able to save between $3,151 and $3,619 on interest fees[5]. Additionally, he will not be required to pay the $25,000 raised during his fundraising campaign, yielding a total savings between $28,151 and $28,619. It is evident that our system has the ability to save individuals significant amounts of money to reduce their financial burden ultimately increasing the chances of successful endeavors. It will also as the ability to play a vital role in reducing the aggregate risk profile of all lending club loans.

Results and Discussion

Of the aforementioned Lending Club models attempted, Linear regression performed the worst and the others performed similarly, with the ensemble model winning by a small amount. The ROC curves in Figure 4 shows the performance of each of the 7 “weak” binary classifiers from the ensemble. While this visual is useful, note that it does not take into account ordinality and the “distance” of incorrect errors. We will show distance based errors in a later table. To combine the output of each of 7 “weak” binary classifiers, we attempted to employ Error Correcting Output Coding. However, since each classifier returned a real valued probabilistic output, we found is difficult to find the optimal cut-off

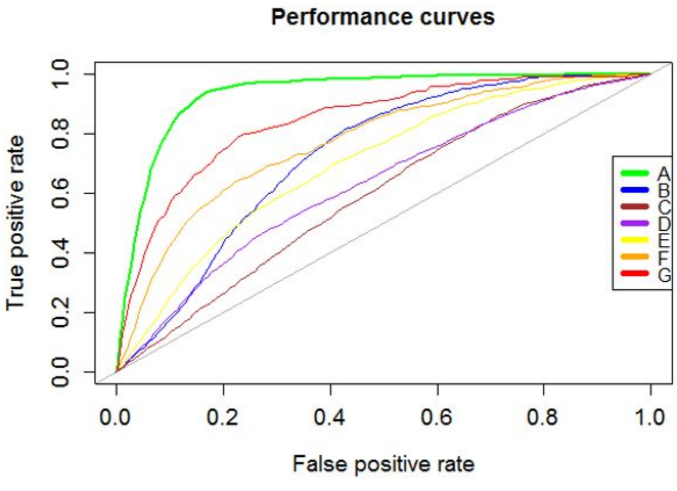


Fig 4

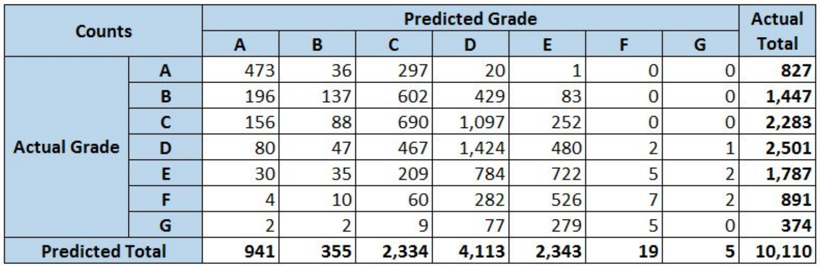
[4] LendingClub loans are originated with a term of 3 or 5 years.

[5] Assuming the borrower does not make any payments throughout the term of the loan.

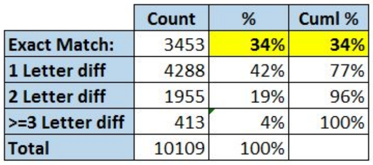
scores for all 7 models. Therefore, we instead employed Maximum Margin Error Coding, which yielded much better prediction results (shown in Table 2). Max Margin simply chooses the winning classifier by choosing the model with the highest output value, which could also be interpreted as confidence. (Allwein, 2000).

To handle the ordinal nature of the rating class, we tried a few techniques. One was to map the A-G values to 1-7 and using linear regression. Another was to queue the voting from the best to worst performing models (compare “A vs not A”, “A or B” vs. “not A nor B”, etc). Unfortunately, handling the ordinality while still employing the Max Margin approach was not compatible, so ultimately we decided to keep using Max Margin. However, we were still able to measure the model based on a simple distance metric between letter grades. Table 2 shows that although only 34% of the risk ratings were predicted exactly correct during cross-validation, another 42% were only off by 1 letter (e.g. model predicted “D” when actual grade was “C”). For comparison, the default model (most common class, “D”) predicted only 24.7% exactly correct, and 40% off by 1 letter.

Table 2. Actual vs. Predicted Counts of Risk Grade Rating.



Counts		Predicted Grade							Actual Total
		A	B	C	D	E	F	G	
Actual Grade	A	473	36	297	20	1	0	0	**827**
	B	196	137	602	429	83	0	0	**1,447**
	C	156	88	690	1,097	252	0	0	**2,283**
	D	80	47	467	1,424	480	2	1	**2,501**
	E	30	35	209	784	722	5	2	**1,787**
	F	4	10	60	282	526	7	2	**891**
	G	2	2	9	77	279	5	0	**374**
Predicted Total		**941**	**355**	**2,334**	**4,113**	**2,343**	**19**	**5**	**10,110**



	Count	%	Cuml %
Exact Match:	3453	**34%**	**34%**
1 Letter diff	4288	42%	77%
2 Letter diff	1955	19%	96%
>=3 Letter diff	413	4%	100%
Total	10109	100%	

One of the hypotheses that our project relied upon is validated in the Figure 5. We hypothesized that the loan amount requested for Lending Club would adversely affect the favorability of the risk rating. For example, the graph shows that as the loan amount increases, the distribution of the most favorable rating, “A” in green, decreases and declines to almost 0% incidence above $60,000 loan amount, while less favorable “E” rating in blue increases dramatically.

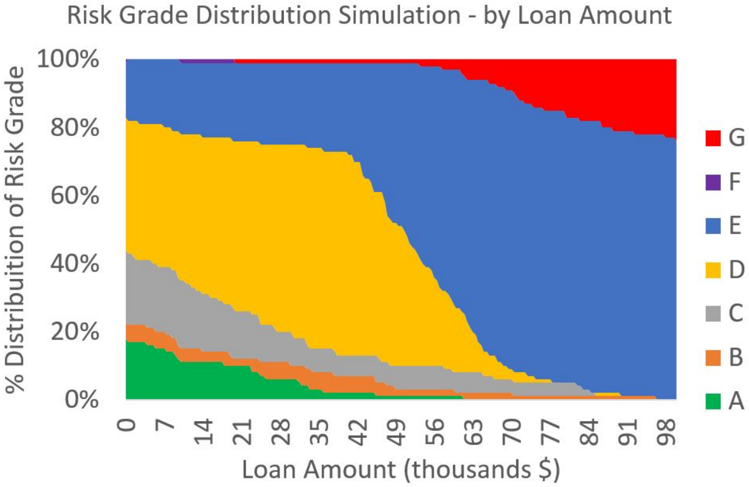


Fig 5

Our second model was used to predict the probability of funding a successful Kickstarter campaign. The Adaboost model slightly outperformed logistic regression for the same feature set while Naive-Bayes fell far behind both models. The former had the highest level of accuracy

(66.60%) while maintaining the lowest false positive rate (21.74%) while the second achieved an accuracy of 64.74% and false positive rate of 22.56%[6]. Even though the accuracy of our two best models only outperformed a default model making predictions based on the prior incidence of successful or failed campaigns by 8.53% and 6.67% respectively, it is particularly important to place an emphasis on the model that can achieve a low false positive rate as we believe it is more important to ensure that we do not tell the user that he will be able to successfully reach an unrealistic fundraising goal. Figure 6 compares the performance of our top two models, the AUC of our Adaboost model is .707 and logistic regression .689, indicating that Adaboost is slightly superior in distinguishing successful from failed campaigns.

It is not surprising that Adaboost outperformed our logistic regression model due to its ability to allow many weak learners to specialize on different subsets of data (Polikar, 2006). Many weak learners are likely to be trained on the records that are harder to classify. In this way, difficult records are attempted to be classified many times to increase the probability of a successful classification (Polikar, 2006). This algorithm is effective because each record is reweighted after each weak learner is trained. Each weak learner has a very low accuracy; however, when used together, they outperform many modeling techniques. Naive bayes and logistic regression do not attempt to treat each record differently nor do they build a model ensemble within the algorithm. Each record is treated equally no matter how difficult the record is to classify.

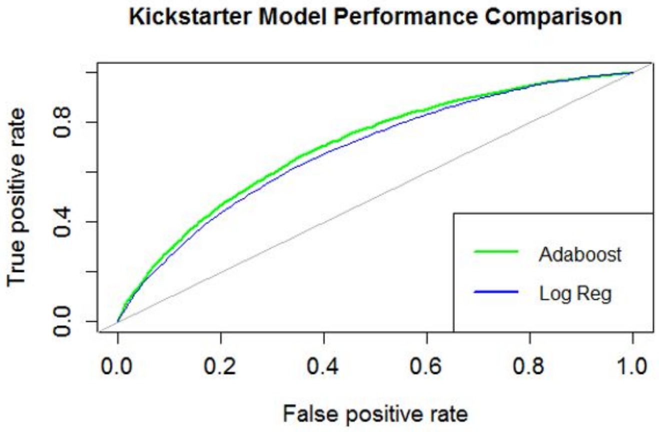


Fig 6

The Adaboost model's concentration on hard to classify records is not the only reason for its superiority. Logistic regression attempts to distinguish between two classes by drawing linear decision boundaries. If there is a lack of linear separability in the output classes (even when including significant features) the model will tend to underperform. Naive bayes attempts to maximize the posterior probability by combining conditional and prior probabilities together. It is likely that our data was too messy to take full advantage of this useful technique. Kickstarter campaigns originated from over 800 states and regions, many of which are duplicates with spelling errors. This messy column could have been negatively impacting the performance of our model by increasing the prior probability and thus positively impacting the posterior, resulting in more successful campaign predictions. This hypothesis is validated by a higher true positive rate (67.20%) and false positive rate (44.94%) than our Adaboost model which has a true positive rate of 50.43% and a false positive rate of 21.74%. We attempted to mitigate this problem by using only fundraising campaigns with a distinguishable US. state but realized a 9% decrease in accuracy and a 70% increase in false positives as a result. A serious data cleaning effort could improve

[6] All results reported are using 10-fold cross validation to prevent against overfitting.

our naive bayes model performance but would have to be weighed against the extensive cost of cleaning over 100,000 records.

Risks and Risk Mitigation Strategies

Text Features Increase Risk of Overfitting

Both the Kickstarter and Lending Club datasets had free-form text fields and we believed that tokenizing this text into uni/bi-grams could provide additional features to improve our model performance, to the extent of overfitting. However, even though the bi-grams had local significance, the sparsity of our entire dataset corpus eventually resulted in decreased model performance. We removed the text featurization from our final model. In the future, we could include this analysis in conjunction with automatic/manual tagging of campaigns into our model building schematic if sites like Lending Club or Kickstarter offered this provision.

Limited Common Data Across Various Crowdfunding Platforms

There was a disparity in the dimensions for modeling that were available in the Lending Club data set when compared to the Kickstarter data. Our ensemble model required us to ignore/group certain fields so that they would be compatible across both platforms. For new users wishing to implement our model, getting additional information from them that completely satisfy our data requirements would enable us to provide a more intelligible decision regarding the outcome of each funding platform.

Optimization Parameters Outside the Borrower's Control

Our models are built to inform the user of the likelihood of obtaining funding based on historical patterns. However, since we are working with supervised learning models, the chance of encountering a type of campaign that the model has never seen before is possible. Additionally, there might exist a scenario where the borrower is unable to achieve her desired funding goal even with a very conservative confidence level as the borrower is unable to change any of the significant variables without compromising on the truthfulness of the campaign. Even though our model is built on a very large training data set, in the event that such an association applies, we should consider dropping these data points from our analysis for model reliability.

Future Work

Our project focused on a single organization for each type of platform (lending vs donation.) A direct expansion of our project can be achieved by integrating other portals such as Prosper, GoFundMe, Indiegogo, Kiva and FundRazr to add richer diversity in its applications.

Furthermore, our project was built on aggregated data that severely restricted our analysis. By partnering with crowdfunding platforms and gaining access to more granular data, we can include additional features such as temporal analysis, supporter segmentation and targeted marketing. Technically, this will also enable us to explore building sophisticated kNN models for sites like Lending Club and implementing greater feature selection techniques such as PCA for our predictive models to increase their robustness.

While text analytics is often a useful aid in model building, our text analysis ultimately proved to be detrimental to our model performance. Despite attempting to model various levels of n-grams, and a variety of tokenizers and stemming techniques, the extreme sparsity and noise in the data still remained an issue. In the future, we can propose that these websites add an automatic tagging feature from a more restricted dictionary that will help us filter out the noise and allow techniques such as Naive-Bayes to provide more reasonable results.

Our project was catered towards the recipient, however a system in which all parties benefit is crucial to any crowd-funding initiative. A future endeavour for this project should be to measure the aggregate economic impact of partially funded campaigns to both sides - the investor and the recipient, and how the aggregate risk profile of lending club loans change over time. This provides a two-fold answer to quantifying the risk involved to lending businesses and the risk to investors within these businesses. A lending club investor routinely invests in loans to earn a return on each investment and a one-sided modeling solution could de-motivate such investors.

The long term goal would be to build a self-learning data product that continually updates the training dataset with information from events that actually happen, such as final status of predicted campaigns. Using this posterior information will allow our model to account for heterogeneity in user perception of certain campaigns over time.

References

Alfaro, E., Gamez, M., Garcia, N., & Guo, L. (n.d.). Package 'adabag' Retrieved April 23, 2016, from https://cran.r-project.org/web/packages/adabag/adabag.pdf

Erin L. Allwein, Robert E. Schapire and Yoram Singer.
Reducing multiclass to binary: A unifying approach for margin classifiers.
Journal of Machine Learning Research, 1:113-141, 2000.
http://www.cs.princeton.edu/~schapire/talks/ecoc-icml10.pdf

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Mag. IEEE Circuits and Systems Magazine, 6*(3), 21-45. doi:10.1109/mcas.2006.1688199