NC3

NATIONAL COALITION
for COMMUNITY CAPITAL

CMSE

Local Invesment Impact Analysis

Amber Agusti, Griffin Frendo, Harshan Ramakrishnan, Christopher Wille

Michigan State University

Table of Contents

Abstract..3
Background...4
Methodology...4
Impact of local investing on small/growing businesses...4
Investment Correlations...5
Lasso/Ridge...5
AutoML...5
CatBoost..6
Impact of local investing on small businesses' communities...6
County GDP Correlation 6
Results...7
Impact of local investing on small/growing businesses...7
Investment Correlations...8
Lasso/Ridge...9
AutoML...10
CatBoost..11
Impact of local investing on small businesses' communities...14
County GDP Correlation 14
Concluding Discussion and Future Work..15
Acknowledgements..15
References...15
Appendix...16

Abstract

Local investing, which involves individuals investing in businesses within their community, can have a significant impact on small businesses. This type of investment not only provides much-needed capital but also helps build stronger relationships between businesses and their local community. By investing in local businesses, individuals can help create jobs, support economic growth, and foster a sense of community ownership. Moreover, local investing can be a powerful tool for small businesses looking to expand their reach and attract new customers. However, it's important to note that local investing is not without its risks, and investors should carefully evaluate each opportunity before committing their funds. This paper will analyze all these factors that went into investing in local businesses with the usage of Machine Learning algorithms to analyze these factors.

Background

This course aims to give students the opportunity to use the skills they have learned throughout their degree program. Each student will work with a team to solve real technical problems related to data science.

The National Coalition for Community Capital has kindly sponsored one of these learning opportunities with a unique problem. Following The Jobs Act, which gave non-accredited investors the ability to invest in small businesses, and Michigan's MILE Act, which created investment crowdfunding at the state level, the group wants to know how local economies have been impacted. Upon receiving the data, there was much done to understand what the dataset could answer and how the data can be improved. There are two main relationships that the team aims to demonstrate within the data. On a large scale, the team wants to demonstrate the affects local investing has had in some communities. More specifically, data analysis will be done to determine how local investment has impacted the success of individual projects, including starting and expanding businesses that were crowdfunded via this program. From there, the team will aim to create a model that can predict business growth based on this local investing initiative.

Methodology

KingsCrowd provided the data for this project. In the raw form, the data contained 5837 rows and 283 columns. Prior to the development toward any of the goals, the dataset had to be cleaned and prepared for coding, this is what is called data preprocessing. In this initial step it was realized that the data contained columns that were nearly or completely empty. After many iterations it was determined that it was best to remove columns that were more than 80% empty. After preprocessing, the data contained 209 columns.

Impact of local investing on small/growing businesses

The approach to this goal was to use machine learning to predict the difference between revenue and money raised. This was chosen because it was believed that being able to predict revenue based on money invested based on features in the dataset, such as industry, could be beneficial for NC3.

Investment Correlations

Plotted the relationship between revenue and money invested in the company to see what exactly the correlation was between the two features. This plot was then fitted with a linear regression line to show the general trend of the correlation. This was repeated but with the relationship between money invested in the company and the total assets of the company itself.

Lasso/Ridge

The model building process began by building Ridge Regression and Lasso Regression Machine Learning estimators using the many features in the KingsCrowd dataset. Since there are a lot of features to sort through, it may be useful to try to use a model that minimizes or completely gets rid of a feature's weight in the model when predicting. In Linear Regression, we try to find a line that best fits the data. However, sometimes the model ends up with a line that fits the data too closely, which means it may not work well for new data. Lasso and Ridge Regression are techniques that help us avoid this problem. They work by adjusting the weights given to the different features in our data. In Lasso, some features are ignored completely, so they don't influence the final prediction. In Ridge Regression, all features are considered, but their weights are adjusted to reduce their impact on the final prediction. Overall, Lasso and Ridge Regression help us find a good balance between fitting the data too closely and not fitting it enough. To begin the process, the dataset was split into a training dataset and a testing dataset, which when completed, both models proceeded to be trained on the training data. The models were then fitted on the training dataset's X features and Y target where X are all the features of the dataset except the target feature in question and all features that directly correlate with it while the target is the difference between revenue and money raised. From there the model tried to predict the testing dataset's Y target based off the testing dataset's X features. The performance of both models was measured by plotting the actual target feature versus the predicted feature as well as the observe how the Lasso penalty affected the weights of various features in the Lasso Regression model.

AutoML

PyCaret was used as the AutoML tool of choice. Pycaret performed all the necessary preprocessing for the data, this included numeric imputation, categorical imputation and one hot encoding for categorical features. All features (209 total) were used in the model. PyCaret tested 19 total models and calculated regression evaluation metrics. 10 kfold cross validations were also performed on each model. In the end, CatBoost was the most accurate model.

CatBoost

CatBoost is an open-source gradient boosting machine learning algorithm developed by Yandex. It's designed to handle categorical features and missing values in the input data. CatBoost uses ordered boosting, a variant of gradient boosting, to handle categorical features. It also includes gradient-based imputation to handle missing values. CatBoost has advanced features for handling overfitting, multi-class classification, and large datasets. It's a powerful and flexible algorithm that's competitive with other popular machine learning algorithms and has been used in a wide range of applications. Due to these features and many tests with different numbers of validation, it was determined that CatBoost was the best model to move forward with.

After discovering CatBoost was the most accurate model, the model tuning process began. This process includes creating a model in PyCaret specifically using CatBoost. This base model runs 10 folds with different splits of the data for training and testing to get an average of multiple regression evaluation metrics. After the base model is built the tune_model function is called. This function ran a grid search for the best hyperparameters for the data. After the model was tuned the model was finalized and it is included in the result section.

Impact of local investing on small businesses' communities

Early during the project, it was determined that local community data (such as GDP, job growth, employment rate) could not be acquired. This data does exist, but it was not in the scope of the project timeline to spend time waiting for data that may not be able to be acquired and analyzed in time. The decision was then made to focus on county data, which is available at a national level. This data was acquired from the Bureau of Economic Analysis (BEA). The BEA is a government agency that produces economic statistics for the United States, including measures of gross domestic product (GDP), personal income, and international trade. It's part of the US Department of Commerce and provides critical data for policymakers, businesses, and researchers. The BEA's data is used to inform economic policy decisions, track economic trends, and inform business decisions. Due to the standard and quality of the BEA data, this data was extracted for the BEA website using their extraction tool.

County GDP Correlation

County information was not available as a standard feature in the data provided. To retrieve the county name, a package called geolocator was used to query the counties based on city and state name which were provided in the data, the "raises_city" and "company_state_name" were used specifically. After retrieving the counties for all the companies, some counties that were incorrect, such as 'United Sates', were removed. After these counties were removed, the now NA counties were replaced with the correct county by matching them with current counties that were correct. The final step was to join the GDP data onto the KingsCrowd data, which now included county information, by using the state and city name. The combined dataset contained information from the years 2019, 2020, and 2021. Because individual companies often only had data for two of these years, there was not enough information to compare the same companies over all three years. Thus, all analyses were done including the years 2020 and 2021. To investigate the difference between companies that have been a part of these local investments compared to other companies, GDP projected growth was computed. This was done using the national GDP growth rate for 2021 and GDP of 2020 for each company. To analyze correlation, the difference between investment dollars for these years was plotted relating to the difference of the projected 2021 GDP value with the actual GDP growth in 2021 for the companies. Using the difference for both GDP and investment dollars in the correlation analysis was important to ensure that the whole timeline was equally considered for both variables.

Results

Impact of local investing on small/growing businesses

Some results were represented using scatter plots and fitted linear lines to show relationships between investment variables. Machine learning results which included Ridge/Lasso and AutoML, were also shown using plots representing validation metrics such as learnings curves, validation curves, residuals, feature importance and prediction error.

Investment Correlations

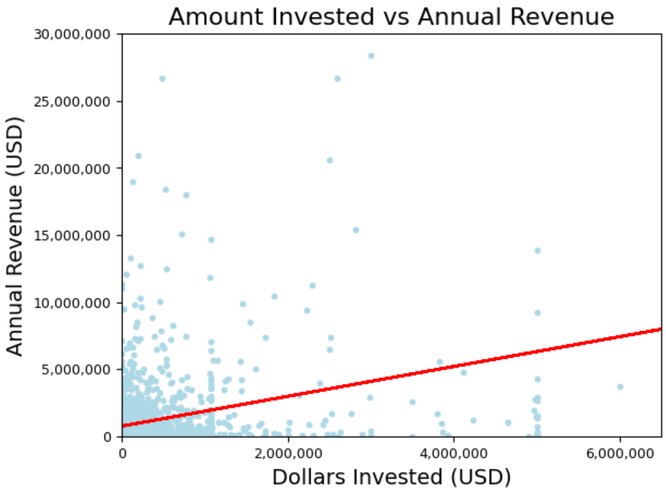

Figure 1. Comparing Dollars Invested to Annual Revenue

Figure 1 showed that as dollars invested increase, so does the companies' annual revenue in US Dollars. This light blue scatter shows the actual data demonstrating this trend as well as a trend line that aims to display this relationship. The equation of the trendline is: y = 1.111 x + 7.49e+05, where the 1.111 value represents the scale increase in annual revenue per dollars invested.

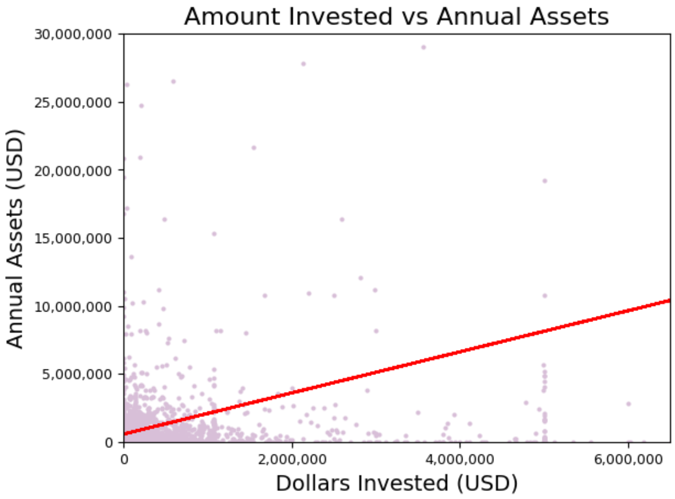

Figure 2. Comparing Dollars Invested to Annual Assets

Figure 2 showed as dollars invested increase, so do the companies' annual assets in USD. This purple scatter shows the actual data demonstrating this trend as well as a trend line that aims to display this relationship. The equation of the trendline is "y = 1.511 x + 5.669e+05", where the 1.511 value represents the scale increase in annual assets per dollars invested.

Lasso/Ridge

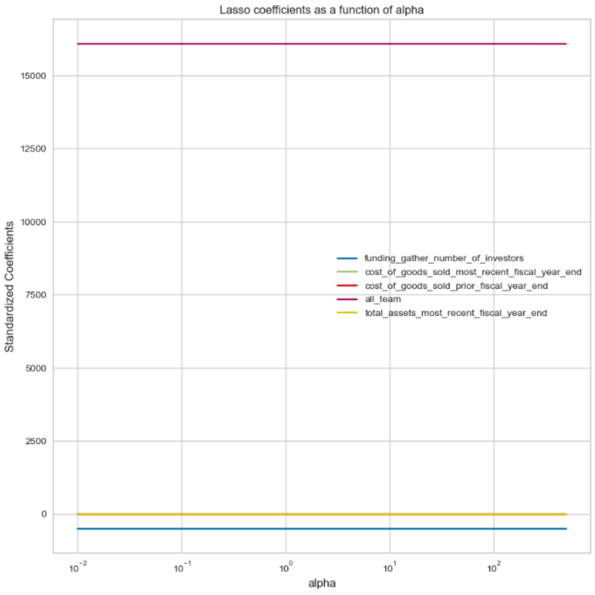

Figure 3. Comparing Lasso Coefficients as a function of Alpha

Figure 3 shows a Lasso regression graph in which the alphas represent the different trend lines of the features used in reference to the standard coefficients when the penalties are applied. In this case the standard correlation is the same no matter the alpha for each of these three features.

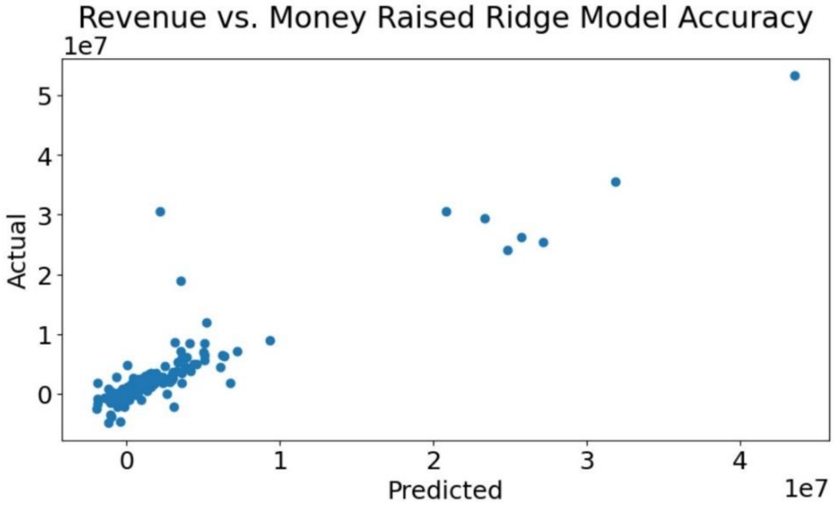

Figure 4. Comparing Ridge Regression predicting the difference between Revenue and Money Raised versus actual difference between Revenue and Money Raised

Figure 4 showed that besides a few outliers, the Ridge Regression prediction got close to predicting the actual difference between the revenue of a company and the money invested in that company. This showed that there are features in the dataset that can assist in predicting this difference, however the model itself did not actually predict any of the values exactly. This led to the search for a different estimator that may have higher performance in the accuracy department. The same plot but for Lasso regression looked identical.

AutoML

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT (Sec)
CatBoost Regressor	479411.2393	5287981170375.5332	2206142.6136	0.5654	3.6408	30.5997	0.8280
Gradient Boosting Regressor	497192.1799	5225900849171.2441	2208103.4205	0.5204	3.8289	31.6981	0.5520
Random Forest Regressor	502393.4113	5594918894475.8750	2286806.2665	0.5181	3.9183	44.7543	0.6010
Light Gradient Boosting Machine	616671.1713	5840155591022.3164	2372089.3000	0.4725	3.8977	53.4566	0.6890
Extra Trees Regressor	451933.7249	6123366825866.0254	2415654.6797	0.4654	3.0408	25.3561	0.5940

Table 1. Top 5 AutoML results

Table 1 represented the AutoML (PyCaret) results for the target variable in regression. MAE (Mean Absolute Error) is the average of the absolute differences between predicted and actual values. It measures the average magnitude of errors in a set of predictions. MSE (Mean Squared Error) is the average of the squared differences between predicted and actual values. It measures the average squared difference between the predicted and actual values. RMSE (Root Mean Squared Error) is the square root of the MSE. It measures the standard deviation of the residuals and provides a better sense of how much the predictions deviate from the actual values. R^2 (Coefficient of Determination) is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. RMSLE (Root Mean Squared Logarithmic Error) is the root mean squared error of the logarithm of the predicted values and actual values. It is used when the target variable has a large range, and it penalizes underestimation more than overestimation. MAPE (Mean Absolute Percentage Error) is the mean absolute percentage difference between predicted and actual values. It measures the average percentage difference between the predicted and actual values. It is shown that CatBoost is the best model AutoML found.

CatBoost

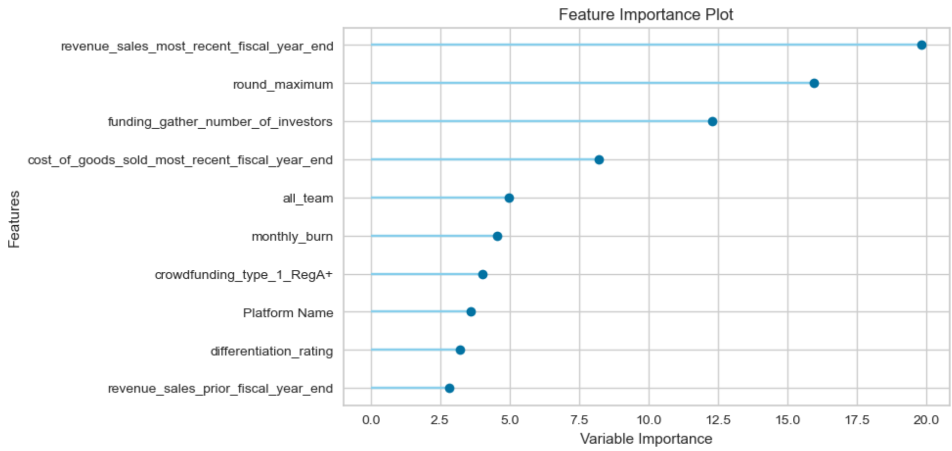

Figure 5. Feature Importance Ratings

Figure 5 shows a feature importance plot. The purpose of feature importance is to show which variable helps the model produce a prediction. As shown above, the model used was the CatBoost Regressor model and the variable in this case which is called feature that was most important for the model prediction was the revenue sales from the most recent fiscal year end.

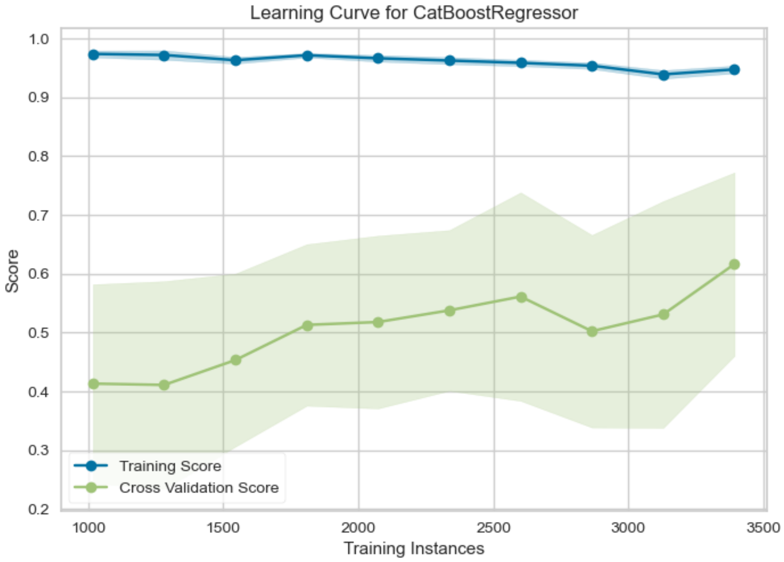

Figure 6. Model Learning Curve

Figure 6 shows a model learning curve. The purpose of a model learning curve is to display the performance of the Catboost model as it is trained and tested with cross validation on more and more instances of data. This figure shows that the training set of the model was predicting quite well with the increase of instances while the testing cross-validation score did increase with the number of instances but was not performing as high.

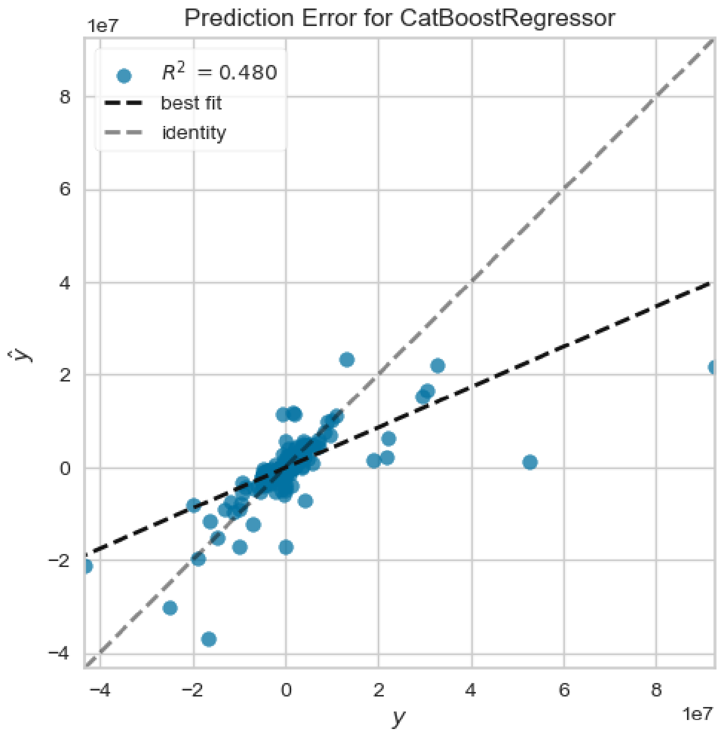

Figure 7. Model Prediction Error

Figure 7 shows a model prediction error plot. The purpose of model prediction error is to show the differing values made from the predicted value of the CatBoost Regressor model versus the actual values. In this case the R-Squared value tells how much percentage of variance there is between the dependent features predicted from the independent features.

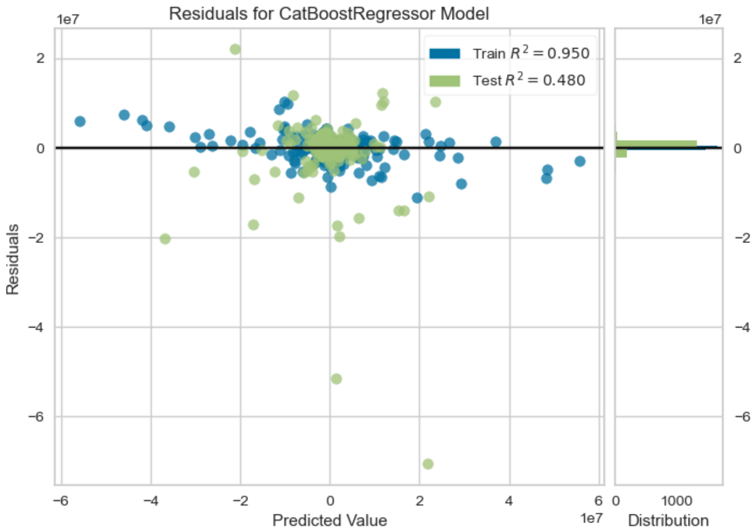

Figure 8. Model Residuals

Figure 8 is meant to show the residual for each predicted value in the training and testing datasets with the idea that most of the predicted points should have a residual close to 0. This seemed to be the case for most points in both datasets except for a few outliers when the model was predicting on the testing data.

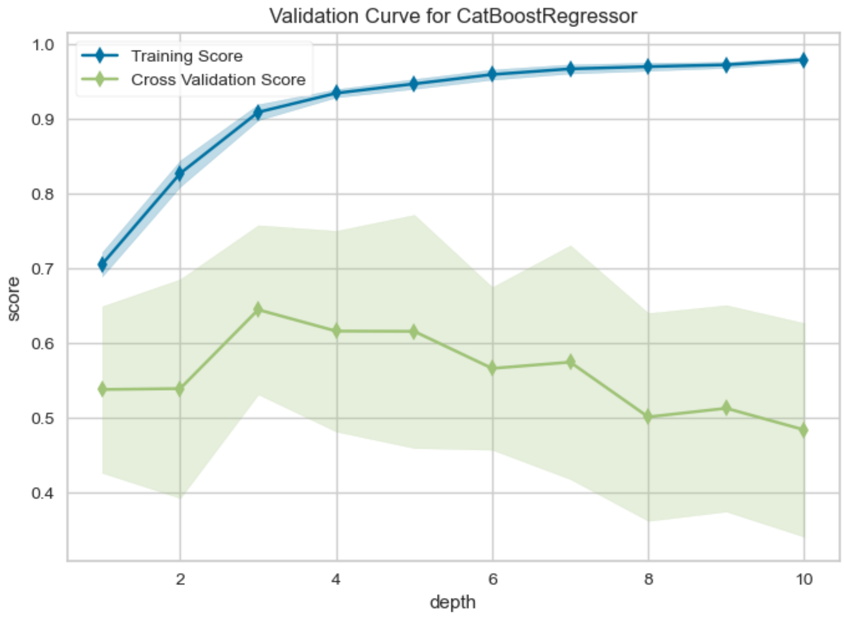

Figure 9. Model Validation Curve

The purpose of Figure 9 is to display the performance of the CatBoost model as it is trained and tested with cross validation as the depth of the CatBoost decision tree is increased. This figure shows that the predicting ability of the model was increasing on the training data as the depth of the tree increased while the testing cross-validation score decreased.

Impact of local investing on small businesses' communities

County GDP Correlation

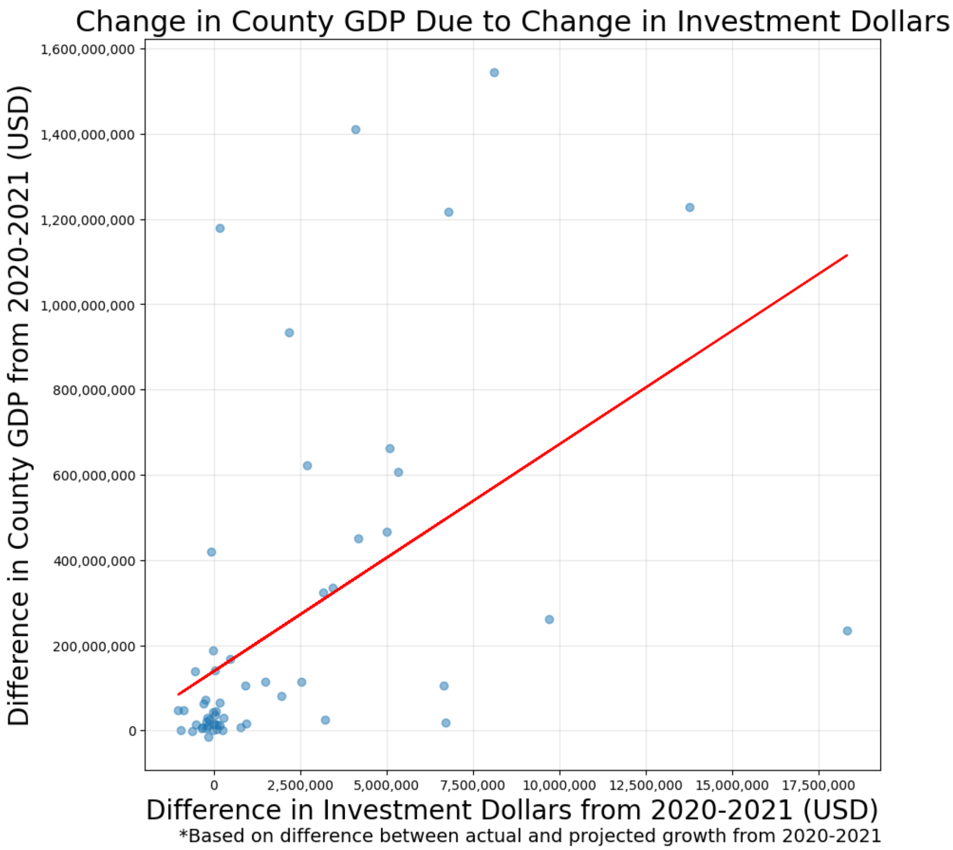

Figure 10. Comparing Investment Dollars to County GDP Growth Between 2020 and 2021

As dollars invested increase, so do the GDPs of the counties that these companies call home. While the x-axis represents the difference between investment dollars between the years 2020 and 2021, the y-axis represents the difference between the projected and actual growth in GDP from 2020 to 2021. A positive value on the y-axis will represent a GDP higher than the projected growth from 2020 to 2021. The use of differences and projected growth was to try to get a better understanding of the businesses involved in this program relative to other, external businesses. Prior to the creation of Figure 10, above, there were less than five outliers removed to better show the spread of the data. The equation of the trendline is: y = 53.27 x + 1.384e+08, where the 53.27 value represents the scale increase in GDP difference per difference in dollars invested.

Concluding Discussion and Future Work

As local investment becomes more prominent, the National Coalition for Community Capital aims to increase awareness and further legislation in this area. To support their projects, this team worked to prove that local investment is helpful to these businesses and their communities. Through visuals in Figures 1, 2, and 10 it is shown that local investment correlated positively to the success of these businesses and the counties that they reside in. To further understanding of local investment and how it affects small businesses, use of machine learning models through AutoML Python packages was employed. The process yielded a CatBoost regression model which was able to somewhat accurately predict a business's revenue to money raised using many fiscal features such as revenue. These figures, models and their implications are helpful to support the National Coalition of Community Capital with their mission to move awareness and legislation towards local investment.

Regarding the future of this project, there are some slight changes we feel would better enhance the results. Providing the future team with possible data from each states' counties, cities and possibly US regions could be better for analyzing and could help investors which industry they would like to invest in for small businesses to potentially make larger impacts.

Acknowledgements

This project would not have been possible without the initiation from the National Coalition for Community Capital. We are especially grateful for the clarification and communication from our sponsor, Mr. Chris Miller, who we hope our results will aid in his and his team's mission. We are grateful to KingsCrowd Capital for their collection and willingness to share their resources with our team. The extensive dataset that they provided us guided and shaped all numerical results. We would especially like to thank Dr. Dirk Colbry who has been a knowledgeable mentor throughout this experience. He has taught us and many others most valuable lessons about data science and professionalism. Nobody has been more important to us than our families whose support carried us through a challenging semester to this triumphant completion.

References

[1] GDP by County, Metro, and other areas. (n.d.). Retrieved March 2nd, 2023,

from https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas

[2] Pycaret Regression. Regression - pycaret 3.0.0 documentation. (n.d.). Retrieved March 24, 2023

from https://pycaret.readthedocs.io/en/stable/api/regression.html

Appendix

Below is a link to the video presentation of this project.

https://youtu.be/Q5E-FeAWUEE

A tableau dashboard is also included in the GitLab repository at ‘NC3\FinalResults\FinalPresentationDashbaord.twbx’ and attached to the final email correspondence.