A Report on Invest NI Funding using Advanced Data Analysis

A report by John Mahon

Email: 21jgm18@queensu.ca

Abstract

The majority of government spending in Northern Ireland is determined by Invest NI, the government body responsible for investing in local businesses. Using data analysis and published data from the Northern Irish Government I will investigate potential trends within the funding of this organisation. I implemented a classification that was able to predict an unknown metric with 89.8% accuracy using the features of the given data. This means that we can analyse if a potential offer of financial assistance will meet our expectations.

Introduction

Invest NI is one of the most important organisations within Northern Ireland, They are responsible for both supporting local businesses so they can compete globally as well as attracting investment from outside Northern Ireland. They do this through a combination of financial grants, business training, and specialist business support.

One of the major recent developments is the end of the Four Year Business strategy implemented by Invest NI [1]. This spans the years 2016-2021, and thus would be a great period of investigation into the investments made by the company, and to compare that to the claimed gains made by the four year plan. To do this, I will be using the published data set of offers of assistance given by Invest NI [2].

This report begins as a general data exploration of a large sample set of data related to something that is important to me. As such, there is not a single question I want to answer with the data, but a variety of smaller ones that we can determine. The provided dataset contains interesting information related to the investments granted, and I intend to use it to answer a few different questions about the investments made by Invest NI, which are explored further in the problem statement.

For these explorations I am going to be using a few different data analysis methods. The data mining process will be easy enough to perform as the API provided gives easy access to the entire dataset. For data processing, there are fields that I need to add for the analysis to be performed. As such, we need to take values from some of the fields and calculate a field from them, writing it back to an appended data file.

Once that is complete, we will perform the analysis on the data. First I will use classification methods from Sklearn to see if we can determine if an investment will be successful based on some of the features of the dataset. For this, we will be using some of the fields that were added during the processing stage.

Problem Statement

As stated before, Invest NI doesn't have a metric in their dataset that explicitly states whether an investment is recorded as a "success" or "failure". For this problem, I want to explore the use of classification models to see if there is a way to predict if an investment will meet a certain grade. The first step of this is to create a number of grades that we can use as metrics of success. Once we have determined a metric, can we use the other features in the dataset to create a predictive model with a high level of accuracy?

Proposed Method

To determine the predictive grade of an investment, I made a pair of grades that have been assigned in the data processing stage of data mining. They are described below:

Average Mean Grade - This grade is used to determine if the return on an investment fell below the average found across the entire dataset. This is calculated by ***1 - (Total Assistance / Total Investment).*** Originally this was just a raw number calculated by finding the difference between the total assistance and the total investment, but was changed because there were strange results which were due to the differing scales of the companies that are contained within the dataset. I plan to calculate the grade and have it assigned to each row in the data, and also calculate the mean value for this grade across the entire dataset. Using this, I would then create a field that had the grade as true or false: true to represent it was above the mean average; or false to show that it was below.

One of the notable aspects of the data is that there were a large number of results that had a grade of 0, which notes that they didn't make any gain in investment from the assistance given. In the chance that these outliers affect the accuracy of the models being created, I also want to tune a model that doesn't include these data points.

Business Plan Grade - This grade is derived from a marketing metric used in the "Four Year Business Strategy" article [1]. In that article, Invest NI uses the metric that for every £1 of assistance they have provided, £6 was invested into the economy. I decided to use this as the second metric to determine the "success" of an investment. This was simple enough to generate, as we just had to perform the following check ***(Total Assistance * 6) < Total Investment***.

Any of the values that we find are true for that check, then we set them to true to show that they have passed this business plan metric. If they fail that check, they are instead set to false in the data to show that they didn't meet this published metric.

Both of these metrics have their weaknesses. Average Mean Grade works off the assumption that only a certain level of investment can be considered a "success", when in actuality even investments below our average can be considered successes. Furthermore, data can be skewed by smaller businesses that have a rate of interest and a higher grade because of the small amount of assistance they were provided. As an example, if a small local business was **provided £1,000 in assistance** and then **invested a total of £10,000**, they would attain a **grade score of 0.9**. However, if a larger external corporation was **provided £20,000 in assistance** and then **invested a total of £100,000**, they would attain a **grade score of 0.8**.In this case, the larger company invested a level of capital just under

nine times that of the smaller business. However, it does show that the company with the higher grade score could be considered more “efficient” with its capital given. In future studies on this topic it would be interesting to see if there were a way to account for this, perhaps with the advice of a commerce student.

The Business Plan grade is simpler to understand, and if the business has passed a success in this metric it means they also have one in the Average Mean Metric. This one is more useful because it is what is used by Invest NI in a review of the investments they have made. There are some odd instances when the business can pass their business plan grade but not the average mean, and this is when the business has been noted as making an investment but wasn’t provided any assistance.

As noted in our introduction, the dataset I used didn’t have any provided metric for success so I had to create these two instead. Of the two, I think that the business plan grade is of greater utility as we can use it to benchmark companies without bias. The Average Mean Grade has the problem that if all investments are doing poorly, then there is a higher level of “success rate”, whereas the business plan is static. In future studies, this could be replaced with a more nuanced investment metric such as the Sharpe Ratio.

Now that the metrics that we are trying to predict have been decided, I will now go through the features that will be used in our classification.

SME stands for Small to Medium Enterprise and is used to show if a business has a number of employees above or below 250. I believe that this feature is important as a way to classify our grades, as the size of a business should have an effect on their ability to invest back into the economy.

Ownership shows if the owner of the business is local to Northern Ireland or if they are externally owned.

Jobs Created is a field that we created for this dataset, and is used to check if jobs are being created as part of this investment. The number of jobs isn’t checked, it is just a check for if any are being created as part of the conditions of the provided assistance. This might affect the investment because of the expectation that jobs are created as part of the process.

Sector is the sector that the business operates within. This could affect the investment compared to the assistance, for example investments into Research and Development could be expected to have a lower level of investment on top of the assistance provided. Thus, the grade could be more likely to fail because of the industry given.

Condition is the condition that the assistance has been approved for. In recent years, as Covid has damaged the economy, Invest NI has been providing additional support and this is noted as one of the conditions. It is more likely to not meet the various metrics if the condition is for Covid Support.

I will use these features and the metrics within the classification model. There are a few steps that we need to perform so that we can train and test the model, the first is that we will take a section of our data that will be used as the hold-out data that we will use for the final testing later on. The data that isn’t included in this will be used for training the model.

This training data is then split again, the trainer which contains every feature except the one we are trying to predict, and the tester that only contains the value that we are trying to predict. We then list out the categorical features, those that aren’t numerically based, and then create a Sklearn pipeline using an imputer and a one hot encoder. This is executed

using a transformer that applies the changes that the pipeline makes upon each of the listed categorical features.

The output of this transformer is then fed into a regressor that uses a XGBoosted classifier. This uses a binary logistic objective, which means that it will perform regression and classify it into one or two classifications. This means that using the provided features, the regressor will filter the prediction into one of two values. We then create a group of parameters for the regressor, which includes the number of estimators used as well as the max depth of the XGBoost tree used. I then finally create the Grid Search classifier using these pipelines and parameters, with the scoring metric “accuracy” being used to determine how effective our classifier is. We then fit the classifier to the trainer and tester data. Once we have completed this we can test the model by having it predict the values for our held-out data and seeing how accurate it is.

Implementation of Solution

The solution has been programmed in Python, the repository of which can be found in the references section of this report [3].
I followed the steps that were laid out in the proposed method. The first step was retrieving the data and then applying the tags that were needed for the predictions. First, I used Python’s Requests library to retrieve the data from the Open Data API service called CKAN. I then proceeded to add the tags which included the business plan grade, mean average grade, and the job created. This involved reading the data into a list that was then updated and written into a new appended data csv. During this step, I also created the zero removed data file, which had stripped out the entries that had not made any investment gain from the provided assistance.

This process was easy to implement, the only issue being that CKAN had a preset that limited the amount of rows that could be retrieved from the API was 100, but this was solved by editing the request to the API. I also changed some of the headings to be more succinct, for example having “Country” as a heading as opposed to the verbose “Country of Ownership when the offer was made”.

While the above processes were completed, I also decided to have the script print out information related to the mean grade averages, both with and without the zero gain entries included. This has been included in the experimental results section below.

With the data mining process completed, I then moved onto the creation of the classifiers that will attempt to predict if an investment will meet the mean average or the business plan grades. For this, we read in the data files that we had created previously into the script and kept the first 6500 and 6000 data entries for the full dataset and the zero removed dataset respectively for their testing datasets. These numbers were chosen so that there was around a 80-20 split of testing data and holdout data.

For the creation of the classifiers, I made three different functions, one for each prediction we would like to make about the data which would then return the classifier when the function was called. Each classifier was created as described in the proposed solution section. Once they were created, I used the *GridSearchCV* built in function to get the best score from each of the classifiers and output them into a terminal so that we could compare them. These results have been given in the Experimental Results section below.

The final step was to then test these classifiers against the hold-out data we created at the start of the classification process. For this, they were called and used to predict the entries that were left out and then write the results to a csv file. These steps were done for each of the three classifiers and a test file was created as a result of each one. These were then compared to the actual results in our original csv files and a score was calculated using accuracy as the measurement (***TP + TN / FP + FN + TP + TN***). These results have also been output to the Experimental Results section below.

Experimental Results

The first set of results we are going to look at are the mean averages for both the full dataset and zero removed dataset.

```
Grades are a ratio of 1 - (Total Assistance / Total Investment)
Average Investment Grade Across 8475 investments: 0.5802254626222648

Average Investment Grade not including Zero-Gainers Across 7739 investments: 0.6354064860736135
```

Here we can see that the average gain grade is 0.580 for the entire dataset, which means that for every £1 invested by Invest NI over the last 5 years, on average £2.38 has been invested back into the economy of Northern Ireland. If we account for those investments that had a return of 0 and thus heavily skew the data, we instead of a grade of 0.635, and this value means that for every £1 of provided assistance, there is a total investment of £2.74. This is a slight improvement, but a far cry away from the publicised £1 of assistance gives £6 of return investment. This might be because they don't include certain low return investments, but I believe it is beyond the scope of this report to try and find the answer to what is considered an investment for this metric.

Moving onto the classifiers, we can see the best score found for each of them below, using accuracy as the scoring system.

```
Average Grade Grid Search
Best Score: 0.7469230769230769


Non-Zero Grade Grid Search
Best Score: 0.7689999999999999


Business Grade Grid Search
Best Score: 0.8979999999999999
```

Here we can see the best scores that we got for each grid search. The numbers show how often our classifier correctly predicted the missing feature, 0.747 means that it was correct 74.7% of the time. We can see that the Average Grade gave us the worst classification result

using these features, with a slight improvement for the non-zero average grade classification score. This means that our parameters may not have been the best for these classifications.

However, we see a massive improvement when it comes to the business grade classifier. It is able to predict if a business meets the business plan grade 89.8% using this test data, which shows that there is a strong connection between the features and the missing business grade value.

While seeing this result for the business plan grade is a good indicator, I need to make sure that this isn't a result of me overfitting the data to this specific dataset. To check this, I then completed tests on the holdout data using these three classifiers for the predictions.

```
Testing Mean Average Grades:
Success Rate: 0.7336708860759493


Testing Zero Removed Mean Average Grades:
Success Rate: 0.7734330074755607


Testing Business Grades:
Success Rate: 0.8850632911392405
```

Here it can be seen that there is a similar level of performance on the held out data, which means that our models haven't been overfitted to the training data used. These similarities give me confidence about the scores for these classifiers.

These results are interesting because as stated earlier in the paper, it is harder for a company to pass the Business Plan Grade that we created than it was the Average Mean. This means that those that do pass both of the values have more specific features that the classifier is able to recognise and thus have a greater ability to predict. Another observation is that removing the Zero Gain entries increased the accuracy of our model in both the training and testing phases. This means that for the entire dataset, there is a wide variety of supporting features that could result in there being an investment gain of 0 and removing them meant that we could get more accurate predictions.

The result for the Business Grade is very exciting, as it means that we can with confidence predict if a potential offer of investment will result in a gain of investment for Northern Ireland based on the other features of the company. With an accuracy of between 88% and 90%, this model could be used to advise what potential offers of assistance should be granted.

Conclusion

This report explored the possibility that data classification models could be used to predict if an offer of assistance would meet certain standards that were defined in this paper. I was surprised at the accuracy that one of these models was able to achieve, as this shows that the features given can be used to get a good measure on whether or not Invest NI will get a return on their investment or not. This is an important factor that we can make predictions about, as the company is the main branch of government investing and spending for the

economy. If this model could improve the process with which they determine who to give their offers of assistance to, then I consider it a success.

There are other elements of the data that are unfortunately unexplored, the main one was the effect that Covid had on the spending of Invest NI. When I started this study, I believed that it would be possible to construct a time series with which to analyse the effect that Covid had on the spending of Invest NI. However, the dataset that I used was limited in that it only contained the entire financial year that the offer was made in and thus didn't have the continuous data that would be needed to construct such a time series.

In the future, I would be interested in finding how this data could be obtained and for a more in depth study to be performed. For now however, I will leave the classifiers that have been created along with their tests.

Links and References

[1] - Four Year Business Strategy Final Year - Four year business strategy ends with a year of significant change | Invest Northern Ireland
[2] - Invest NI Offer of Assistance Dataset and Accompanying Guidance
https://www.opendatani.gov.uk/dataset/open-data-up-to-17-18-csv-file-uploaded-csv-13-to-2016-17/resource/cd00d300-fcde-4ad8-921e-f1324b75d37e?inner_span=True
Invest NI Financial Offers of Support 2016-17 to 2020-21 - Invest NI Open Data Guidance
[3] - Python Repository - https://github.com/Jaygeepd/Invest-NI-Data-Anaylsis