# Define the problem

Introduction

Our company calculates legally qualifying expenditure related to R&D projects in order to retrieve tax relief for our clients.

We operate on a success based fee structure as we will take a proportion of the final  
benefit due to our clients. In the course of securing this benefit, we will proceed through an initial call, physical meeting and/or further call, collect financial documents and expenditure data, compile a justification and submit our report to the tax authorities. At any stage, legal, technical or financial factors may hinder the process, or prevent success completely. We operate in a small market, it is a niche industry, and our clients will not have a clear understanding of where our services will take them, and it is likely that we won’t know either due to the complexities of the industry. We need to establish greater clarity on our process as early as possible in our process, so that both our internal staff and our clients understand the eventual benefit or cost of continuing. What would be the potential benefit to our clients who don’t know anything about this specialised industry? We have excel spreadsheets per client and per annum, with data showing how much qualifying expenditure companies have paid relevant to R&D, which is the basis of the tax relief.

# Identify your client

My primary client is Catax, and I want to strengthen the company’s understanding of potential success or failure in a client’s claim for tax relief, and to improve upon the dialogue with our clients by providing effective tools for Catax. With a better understanding at the forefront of our process, we can substantiate greater evidence for the returns of our product, improve our qualifications of our clients, provide a better service altogether, and improve decision making of how our clients should be handled.

# Describe your data set, and how you cleaned/wrangled it

Extraction of staff costs

The qualifying R&D related expenditure is entered into excel spreadsheets before being extracted into our report to HMRC, called cost summaries. These cost summaries are copied into a folder rather than being extracted from the entire internal company network, which would be a costly operation.

A for loop is used to loop over every worksheet in every spreadsheet. A function ‘get data’ is used to iterate over every row in each worksheet which retrieves the data into pandas Data Frames. A function called ‘staff extract generator’ is used, which removes empty cells and applies the squeeze nan function; this reformats the Data Frames. The index includes two data points, ‘staff cost’ to ‘total costs’. Between these data points are all of the R&D related expenditure. However these data points are not always labelled as such, therefore to extract this data they must be wrangled to be systematic, and then extracted. Lambda functions are useful to change the labels. Lambda is able to approximate the cell labels, therefore we can use this approximation as an if function. If the cell contains the pattern we want, such as including ‘staff’ it can be changed to ‘staff costs’. If ‘total staff costs’ then it is changed to ‘s total costs’. All data from staff costs to s total costs are staff related expenditure which are extracted to create a staff expenditure database.

To source the data, the spreadsheet filename and worksheet name is extracted to a column, which will be utilized as a reference point, allowing us to locate the associated client and period in SQL databases.

The data we require is only in the first four columns of the data set, and the spreadsheets may contain notes, therefore only the first four rows and reference points are extracted from each worksheet.

As the spreadsheets we are iterating over contain worksheets that do not include the data set we require, a try and except function is applied, and the script prints the exception in case there are any worksheets that contain our desired data that are not extracted due to errors.

Cleaning

After extracted all data worksheets individually, they are compressed into a single dataframe. We can now begin to give shape to the final dataset. The column names are titled to indicate the data within the created columns. The job titles are refined. A regex is applied which identifies entries in the job titles column with text only. A similar validation is used against the apportionment and total cost. A lambda is applied to both columns which ensures that they are int or float types only. Any string type entries should be removed. As the entries might be entered incorrectly, MAX and MIN functions are applied to extract the cost and apportionment respectively, under the assumption that the greatest figure will be the total cost, the second figure will be the calculation between the total cost and apportionment (not required) and the apportionment is the lower figure. The apportionment must also be between 0.1 and 100 to be considered a percentage of total costs. If there are any nan cells after this process, the rows are removed.

Extracting identifiers to connect to client data

With the data set correctly formatted, and the bulk of data in correct order, the CIF (ID) of the correlating client can be associated with a database of client data. The cif is not always in the filename of cost summary worksheets, therefore the best way to extract the cif is from the source of the spreadsheets, from our internal servers using the ‘client id’ function.

A Data Frame is constructed with file destinations in the internal drive. A merge is completed between the Data Frame of file destinations and the staff costs column with filenames. To ensure that the original index is not lost during the merge operation, it is reindexed, and the job titles is ultimately used as the index. Using the source path from the internal drive, a regex is utilized to extract the correlating cif number. As many clients have three to five-digit numbers in their company names, the regex excludes specific digits I have recognised within company names in the Data Frame. File names might also include the correlating periods including 2013-2017 or expressed as 14-17, therefore these digits have also been excluded.

The year is then added from the worksheet of each spreadsheet. This might include the periods between certain dates or with a year, therefore a regex is used to identify the last two digits to represent the year-ending period.

Final cleaning

The remaining step to clean the data is to remove duplicates, which is applied by sub setting the staff names and worksheet names. The columns we need for the final dataset are selected, the cost, percentage, cif and year. The R&D related costs are excluded as they can be computed later by cost and percentage. With the cost DataFrame obtained, it can be connected to an Excel spreadsheet which includes the client data.

Other cost brackets

The subcontractor databases was completed with key differences to the staff costs. The labels used to define the location of the data required for this data set are set to xsubcontractor and xtotal sub to create unique identifiers for the data. The columns are not selected until after the worksheets are concatenated and an additional column is added. The additional column is needed as the subcontractors have an additional column to detail whether they’re connected to the client or not – if they’re not connected only 65% of their costs are claimable. Consequently, this dataset is more vulnerable to shifted columns, where a column will have a part of another column’s data. Due to this vulnerability in the data set, the max and min functions used to unscramble the data is more important. The consumable and software data sets are created in the same manner as the subcontraction costs to ensure consistency in the code, but with corresponding labels to extract relevant data.

# Explain your initial findings

Costs brackets

The most important variables are the expenditures themselves, which construct the total costs. Counting by number of clients and periods, there is significant variation in the occurrence of each type of cost. The staff cost database is significantly greater than software expenditure, including 1900 clients and periods of the 2100 samples, while the software database includes 240 of clients and periods. The difference in size of data sets might reflect why the staff correlation coefficient would be much greater than subcontractor, material and software by a large margin. However we have tested by each cost type where they are present, therefore this should not disrupt the results too much. It also makes sense from a project perspective that on the whole, the bulk of expenditure would be staff and the time they’ve spent on the projects, while the materials they use would not be so significant.

There is significant variance within the costs data, with mostly small figures and a few large ‘outliers’; the data is exponentially distributed. It would seem that the majority of our clients are completing ‘side projects’ and a few companies are generating a lot of R&D expenditure. The distribution of each cost total vs number of line of costs indicates that it could be a good predictor that the more of each cost there is, the greater the cost, although this isn’t the case with the material and software costs, with some low correlation coefficients. The material and software costs actually seem to share the same distribution which is proven to be statistically significant. They also have similar behaviours, a similar allocation to R&D towards 100%, and roughly the same number of material/software per client. These behaviours lead to a discussion of how well the descriptive characteristics of expenditure could be used to predict total expenditure, rather than using ‘external’ variables to explain variability.

For example, can cost apportionments predict total expenditure from this? Using the median per client/ period (as the data is exponentially distributed), this metric has not performed well in predicting total expenditure. Staff apportionments had the strongest correlation but it was still very weak. Although as a single variable, apportionments do not perform well, I would still surmise that it would be useful along with the number of staff or payroll to estimate total expenditure. Another characteristic of the costs would simply be counting each cost in each bracket per client and year. Again, it is the staff members that have the strongest correlation of 0.62. The other three cost brackets have moderate correlation to the total costs in this method. Subcontractors have a correlation of 0.39, consumables have 0.22, and software have 0.30. The count of each cost should be an useful contribution to the total expenditure predictions.

Client data

Industry was a feature that I expected would be significant to predicting R&D value – I would expect that a client in the mechanical engineering industry would have much greater R&D expenditure than a client in the food and drink industry, but the difference is minimal, and there is very little correlation to total expenditure. Although some industries are much greater than others in terms of R&D, and certainly in the number of our clients per industry, each claim for tax relief cannot be differentiated by industry. Each of the top four most frequent industry have very similar distributions and it is some outliers that drive differences in central tendencies.

Postcodes was another feature I expected to be useful to predict R&D expenditure. This had a very low correlation score also, and was not useful to predict expenditure. The postcode in particular shows the weaknesses of the dataset as there are so few records per post code. Of the top 15 most frequent post codes, there are only 643 records. Using the post codes, there are some regions with very high values but the correlation was not statistically significant. There would be a similar issue using SIC code data which classifies business activities.

The final variable which I expected to correlate was the phase, the number of times a client has completed the process with Catax. Judging from the mean of each phase, the more times a client works with us the greater the value of the benefit. However, from looking at the data statistically, there is no significant difference, and there is nowhere near as much data to utilize in phase 3 through 5 to make a significant conclusion.

A variable I did not have high expectations for but was curious to see the result was the month end variable. Perhaps there is a difference between companies who choose a specific accounting month ending period. The findings show that there is not a significant statistical difference, although there is a greater central tendency in months such as March. There is also the turnover figure from the client data which could reflect how successful a company is. This is a very small data set, and also at best has a moderate correlation to total R&D costs. I did not expect this data set to be a useful predictor as it is generally only obtained at the point of sale, but evidently it is an effective predictor that should be useful when predicting future tax relief for the client. We should seek this data on an annual period to refine better results, and possibly look deeper into accounting data to search for predictors.

# Deeper Analysis – Machine learning

Regression models

In the most ideal situation, we would be able to predict exactly how much R&D expenditure value there would be based on the client’s data. With this goal in mind, I first began to analyse the data according to regression models, if there would be linear relationships. The most important feature according to the inferential statistics analysis was the staff costs, after calculating from the staff payroll and time spent. The regression OLS function was able to show some useful statistics using the staff costs. This feature had an R-squared value of 0.71, which means how much variability can be explained in the dependent by the independent variable. This is a good performance for a single feature, and it returns a statistically significant P value of 0.00.

There is evidence that this variable would be a powerful predictor as has been shown previously, however I would aim to construct a model without it. Without this variable we can use a model before a prospective client needs to expend too much effort in the process. Accordingly, I first analysed a linear regression model and cross validate the results of all other variables previously described. The results were very poor. We already know that using all of the features, the data is not linear or normally distributed. A CV test can be used to split the data several times and use a proportion of training data to predict the other proportion, to see the prediction performance. The performance of the linear regression algorithm has an average 5-fold score of -329531. A random forest regressor is a preferable choice. This ensemble method utilizes a pre-defined number of decision trees to explore probabilities. Selecting 20 decision trees to get a sizeable number of predictions, we get a much better score of 0.66. testing different number of cross validations yields roughly the same results of 0.63 – 0.66. A quick check of the predictions of the first 5 values of the data without splitting the data is interesting to see the variation and difference between the prediction and the actual data first hand. The predictions are each several thousand pounds away from the actual data. I then used the same OLS regressor function to see the data results. The data has a high 0.839 R squared value, a P value of 0 of the whole model (so the model is statistically significant). Specific features also have very high P values, including 0.97, showing that there are features which are not statistically significant at all, and need to be removed from the model.

On this basis I removed individual variables which were the least statistically significant. I removed CIF employees and phases. They were features with high P values and few categorical values. The postcode and industries P values may change with the removal of the higher P values. With this model, a 5 fold cv test returns a value of 0.67, which is a very marginal improvement upon the previous model. A 3 and 10 fold cv test returns a score of 0.64, which again is not a significant improvement upon the previous test. The high p value features have been removed but the r squared score is maintained at 0.839 and the model’s statistical significance has a p value of 0.00. We have maintained the same strong evaluations of the model while marginally improved its predictive capabilities. I then removed any additional features with the intention of removing any that have a p value above 0.05, but the features that remain have a p value of 0.00, except for the number of staff which is 0.02. With only the salient features remaining, I wanted to test the use of the linear regression on the dataset. The CV score is still lower than the random forest model, at 0.55 – 0.57. The random forest model returns CV values of 0.67 and 0.7. Removing the variables with high P values in the original data set has increased the performance of the predictions but marginally. With the appropriate features and optimum model realised, I tested the mean squared error to evaluate the distance to the correct predictive values. The MSE returns 2417623375.63, which is enormous and incoherent. Visualizing the predictive and predicted values shows that the distance between predictive and predicted values are large but the MSE value shows that the model is unacceptable to make predictions as the project intends.

After seeing the results of the regression models, I decided to proceed to classification models. I binned the total costs using Qcut to five equal parts. As the data is exponentially distributed, basing the bins on data values would eschew the predictive power of the model, therefore I based the bins on frequency. In this way we can predict the values of the smaller values while when predicting the larger values, the binned prediction would essentially communicate that value of the total costs would be over a specific threshold. It is more preferable to create more precise predictions of the lower and more numerous total cost values if the claim is worth completing. It is in this area the predictions would create more power i.e. if the claim is over or under £15,000 of R&D expenditure. This would be important for a prospective client to know whether the process is worth their time, and for Catax. If for example the prediction is over £100,000 then a client would know that there will be a sizeable return on investment.

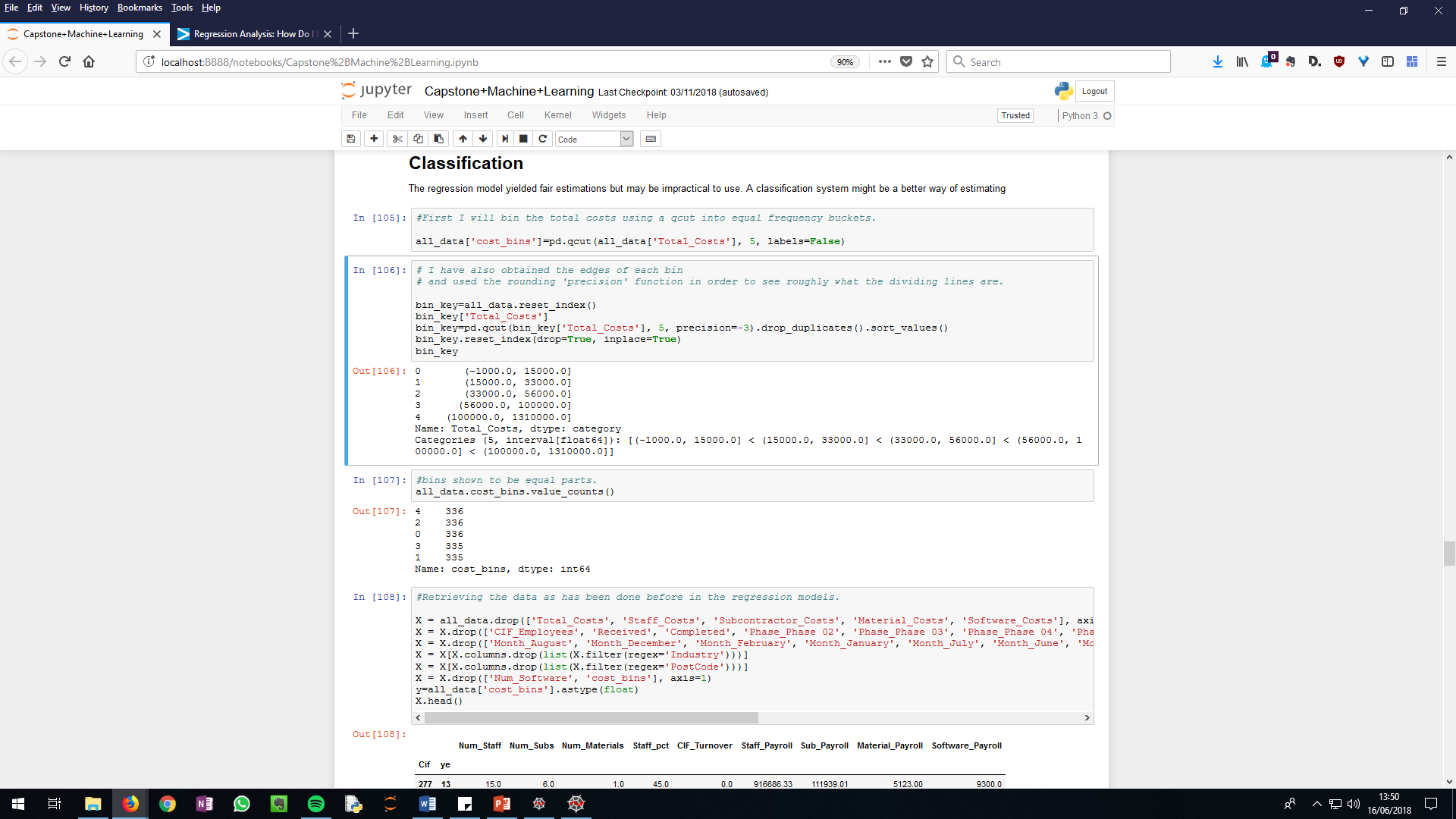
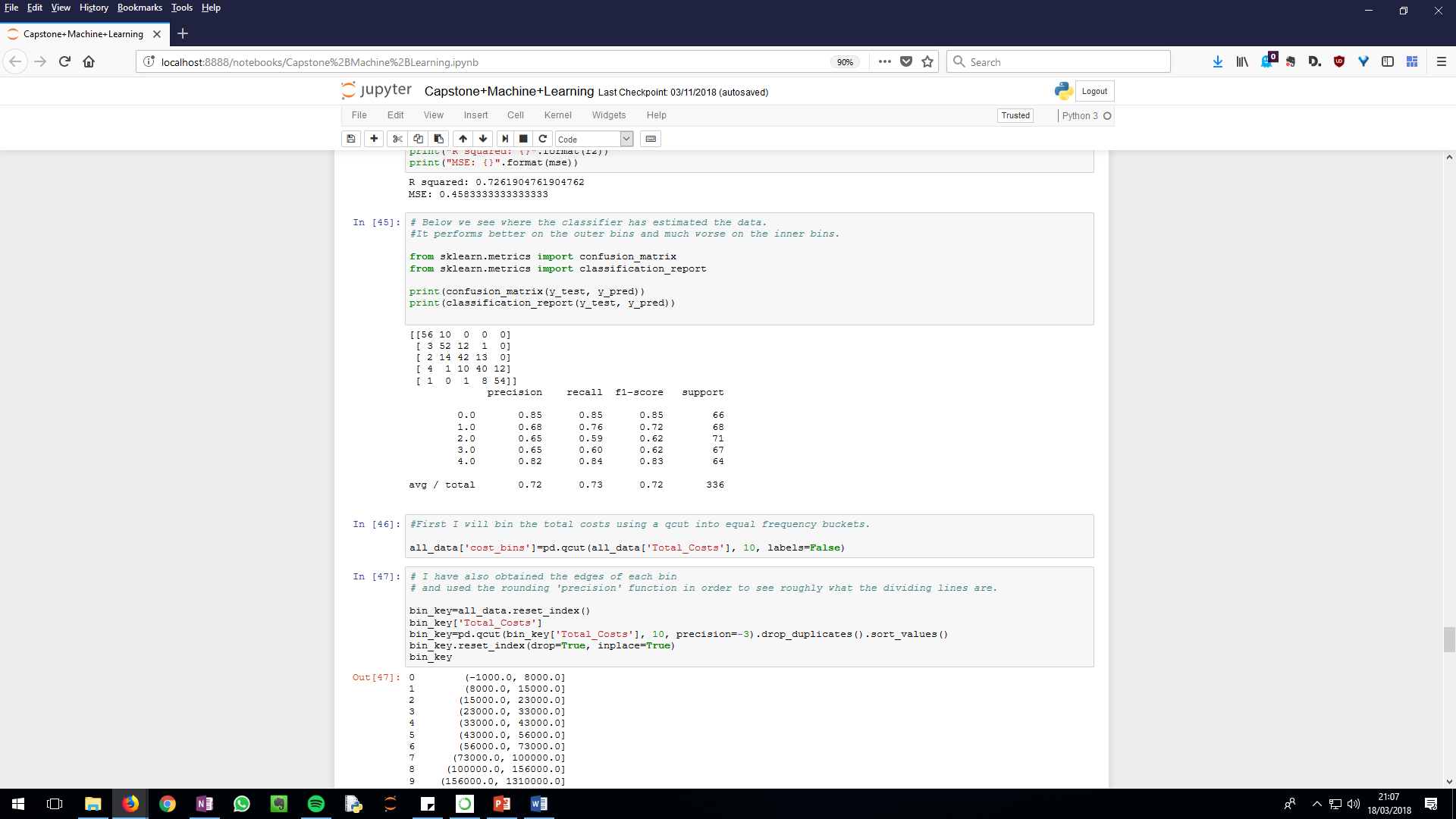
The first classification model is based upon five bins. For visual purposes, the bins are rounded to the nearest thousand in the image on the right.

Figure Grouped total expenditure to the nearest thousand

The regression results from testing this model, on splitting between training and test data has an R squared value of 0.72, which is lower than what would be anticipated, but the mean squared error is 0.45, which is far more manageable in this classification system than a regressor model.

Creating a confusion matrix, we can see how well the model has predicted the values. The predictions are stronger the lower or higher the predicted class. The closer to the centre of the data, the more difficult it is to predict.



Running the classification report function shows how well the classification model has performed, and puts a figure to the description based upon the visualization above. The precision, recall and f1 scores are worse towards the second bin, but are highest in the first bin.

But perhaps it would be more manageable in a practical setting if there were more classification groups. The classification groups were then doubled to 10. This returns a much lower R squared value of 0.49 and a higher MSE of 1.5, meaning there is more error. The cross validation score too is low, being 0.39 and 0.42 respectively with 3 and 10 folds. The classification report returns an average value of 0.5 between each metric. This model is weaker, but it has greater capacity to be more precise. Keeping the results in mind, it would be more effective to judge 0, 1 and 9 classifications to be correct, and between 2-8 to consider the classification to be between ranges. I.e. a classification 4 to be valued between 3-5. A classification 5 to be between 4-6. The classification works in so much as the correct value is the most frequently predicted, along with the ‘edge’ of the respective bin. A classification 5 has 27 occurrences between 4-6 classification bins and 7 out of this range. The problem with a smaller number of classification bins is that we would not know how close a predicted value is to the edge of each bin, therefore by increasing the number of bins and understanding the predicted values, we can adjust our predictions.

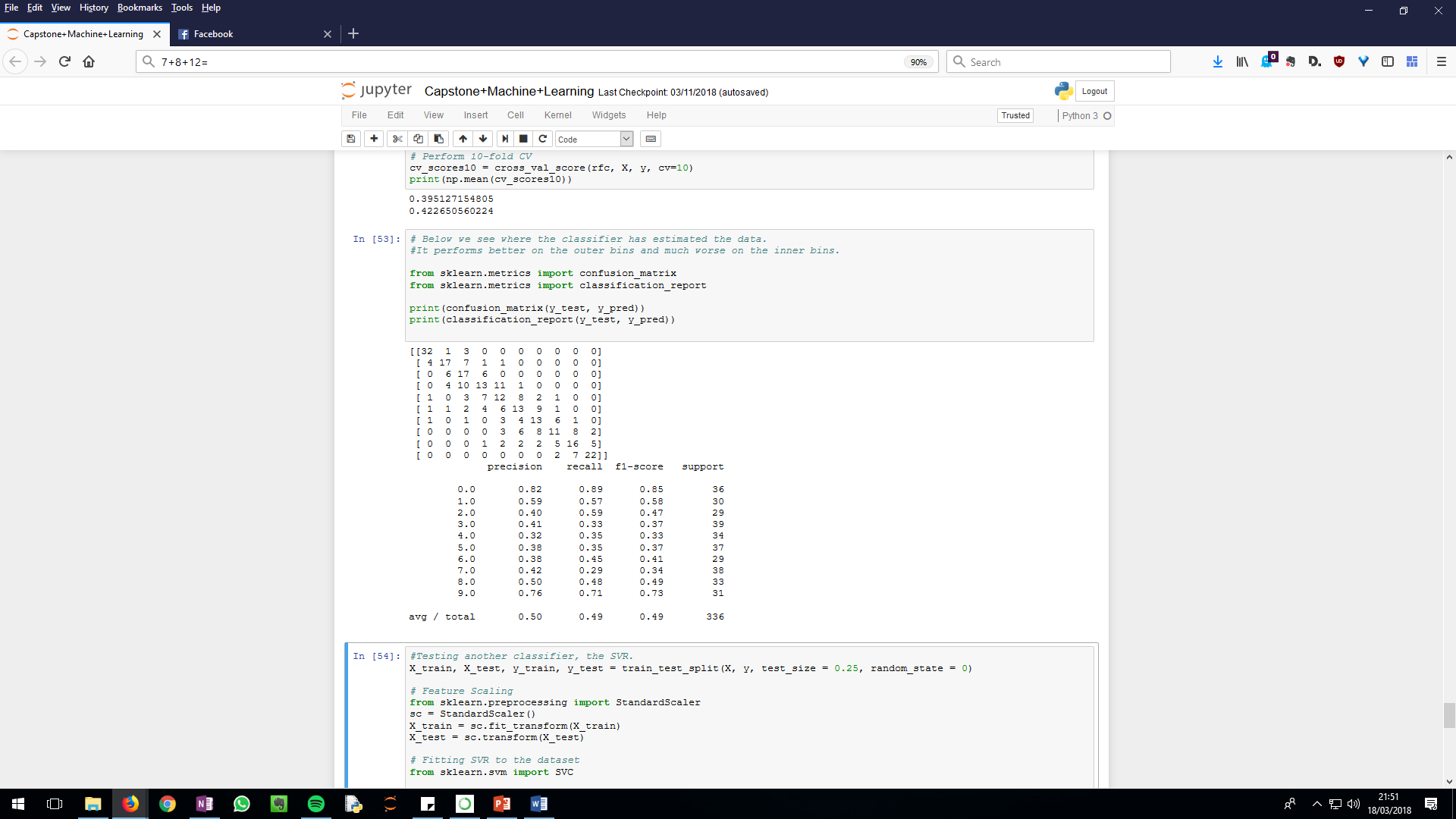


Figure 10 bin RF results

Finally I have tried one other classifier, the RBF kernel. This was chosen as the data is non-linear and by mapping the data to a higher dimension, we could better isolate each class. It is far more inaccurate and predicts across many bins, which is a real weakness if we were to estimate by a range of bins. The random forest classifier is the preferred method.

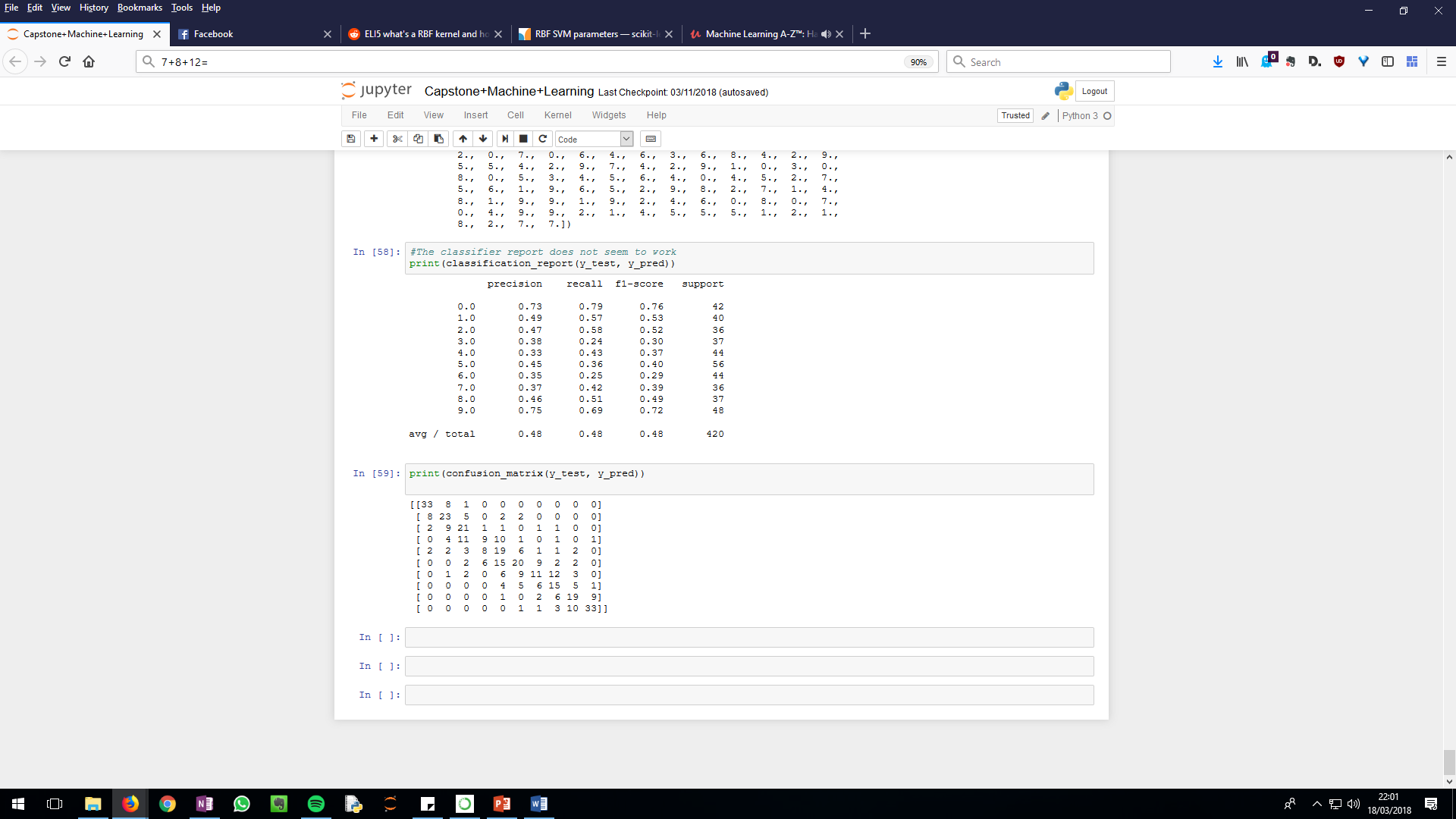
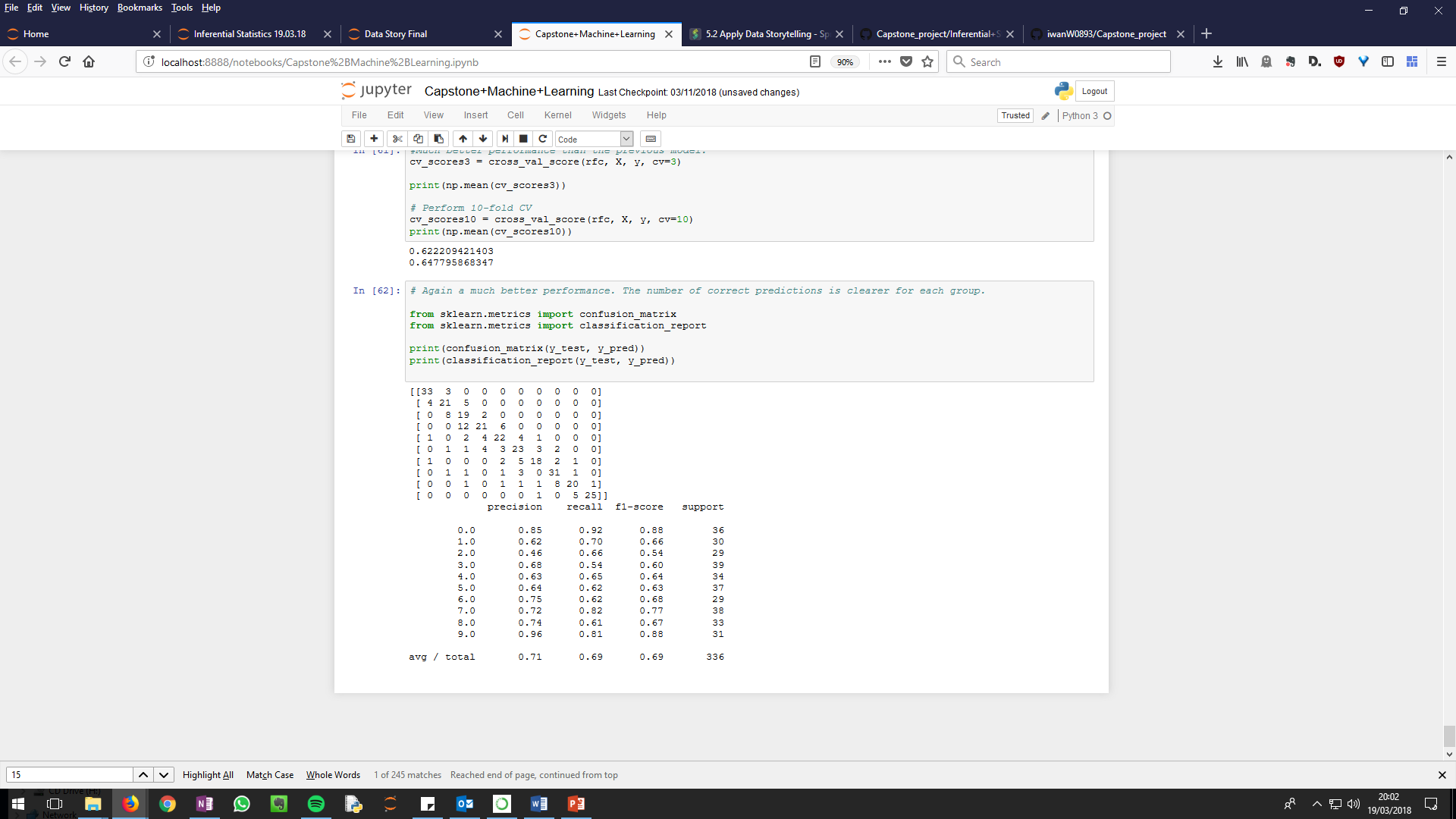


Figure RBF results

With the preferred classifier established, I wanted to try one more model. In the previous models we have been trying to predict R&D expenditure with limited information to improve prospecting capabilities with clients. We can introduce the staff costs, the strongest predictor, and the client could put some thought into how much time their staff members spent on R&D and without factoring other cost brackets. This model would be the next stage of the process and help us to refine the prediction. Predictably the test results are better, the R2 value is higher at 0.69 and the MSE at 1.062. the CV scores yield 0.622 and 0.647 with 3 and 10 fold tests. Per the classification matrix and report below, the model maintains roughly the same accuracy scores as the 5 bin model, although the confusion matrix is much clearer. The predictions are much more successful while maintaining the precision of using 10 bins instead of 5. This model would indeed be more useful further along the process when we can predict using the established staff costs.



# Recommendations and further improvement

Based on the analysis above, I would primarily advise the client to experiment with the models and test their performance at the stage of the process the respective model is used for. The 5 and 10 bin models without the staff costs can be used when prospecting with potential clients. The classifier with the staff costs can be used further into the process when the client has established these costs. As in the past, predictions have been rough estimates, I would recommend testing the models’ performance against the rough estimates. A visualization tool has already been created to calculate the potential benefit to the client, based on their expenditure. By predicting their expenditure with these models, we can combine the two to predict the potential benefit, not just expenditure. But before using these tools to a prospective client they could be used to predict revenue to the company where there is less risk but still an effective use to the business. As the data set the models are based on are small, and without prior planning the data is sporadic and circumstantial it is best to test in the field.

I would also recommend practical considerations to improve the models and this project. As the dataset is very small, the models will improve with more data or data points added to it. We particularly need to pay attention to the ‘CIF’ features, which were only 344 rows out of 1678. The CIF turnover is an useful variable although not the strongest. It should be ensured that CIF turnover is added to the data set continuously with future clients. A notable feature I would further investigate is the industry categorical data. The industry categories are in sporadic sizes and definitions. The number of clients per industry ranges from 382 manufacturing to 22 in the medical industry in just the top 15 most frequent client industries. This should be restructured so that the data is sufficiently broad yet precise. Of course it is important to analyse that for example, manufacturing is an important industry, but looking at the smaller industries, such as funeral planning and pest control services, these are not insightful categories and likely disrupt the predictions. SIC data on the other hand are too numerous and also don’t necessarily relate to the company’s projects. I would recommend breaking up the industries by the projects that our clients complete. This might be a material based project or a software based project, process or product for example. All industries can develop themselves in different ways, therefore this needs to be dissected and analysed for its predictive capabilities. My final recommendation would be to categorise staff involvement on a project basis. We have a data set of what staff roles each staff member had, but if we were to categorise their roles specific to the projects, we might be able to better predict the value of the project. i.e. if the project involved a lot of testing and development this would increase its predicted value, but if it involved a lot of project managers effectively directing its subcontractors to complete the project, the tax relief might not be as lucrative. The models have the ability to predict R&D expenditure to an extent, however I believe that the recommendations above would greatly improve its performance. With more thought put into the idea of predicting R&D expenditure and additional variables and ensuring the data is clean, the models would better predict R&D expenditure.