# 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 for our clients. In the course of securing this benefit, we will proceed through an initial call, physical meeting / call, collect financial documents, expenditure information, compile a justification and submit 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?

# Identify your client

My primary client is my own company, to strengthen the company’s understanding of potential success or failure in a client’s claim for tax relief. 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

Cost summaries are first copied into a folder so that the script isn’t too intensive in scanning over the entire internal drive each time the operation is completed.

Next is a for loop which loops 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 is called, which drops empty cells and applies the squeeze nan function; this reformats the Data Frames to restructure a summarising ‘total costs’ cell to follow a heading ‘staff cost’. Between the heading and total costs in where the data lies. As these ‘labels’ are not consistently applied due to either error or process changes, the labels are modified to be uniform. Whitespace is stripped from the index as a series, and using lambda functions the labels are changed. 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, which translates to a ‘staff’ heading or ‘total staff costs’ then it is changed to s total costs and staff costs. We can then add the edited index back into the Data Frame index and loc can be applied to specify the data required.

The filename and worksheet is then extracted while the spreadsheets are being manipulated, which will be utilized as a reference point, allowing us to locate the associated client in our main systems.

The data we require is only in the first four columns of the data set, and the spreadsheets may contain notes which would corrupt this data set, therefore the first four rows and created 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 do contain our desired data that are not extracted due to errors.

Cleaning

Next is the ‘concatenated df’ function which creates a single Data Frame from each worksheet Data Frame. We can then begin to give shape to the final dataset. The column names are titled, including ‘r&d cost’ which is included in the original set but not included in the final Data Frame; but it is included at this point in case, although unlikely, a spreadsheet included additional information, i.e. the jo role and project role.

A reindex is used so that we can refine the job titles used in the staff costs. A regex is applied which identifies entries in the job titles column with text only, as columns which would have numeric values would class as an entry without a job title or staff name. At earlier stages of our company’s process in treating these costs we have not included the staff member’s name, and in this process these entries are eliminated. Although they are entries that contain apportionments and costs they don’t include the job role, which is vital, and I would consider this data less reliable. 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. Strings are not desired. 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, then the cells are dropped.

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 the database of client data. The cif is not always in the filename, therefore the best way to extract the cif is from the data source, from our internal servers using ‘client id’ function. A Data Frame is constructed with file destinations including costs summary or cost summary as the data was originally extracted. A merge is completed between the Data Frame of file destinations and the staff costs dataset based on filenames. To ensure that the original index is not lost during the merge operation, a reindex is used and the job titles is ultimately used as the index. Based on the foldernames extracted merged into the Data Frame, 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 these recognised digits in the Data Frame. File names might also include the correlating periods including 2013-2017 or expressed as 14-17, therefore these periods have also been excluded. If more than one cif has been extracted from these file destinations, then the function ‘tidy cif’ is used to split the digits extracted into separate rows. The cif column is then cleaned to remove characters created by the lists and remove empty cells, and their associated rows.

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 before the “] characters used in entering the worksheet names, and the ultimate two digits from this created dataset is extracted, ensuring that if two dates are extracted then the final two digits are the digits that represent the year-ending period. At each stage the created series’ have a function applied to remove the list formatting to return only the two digits we require.

Final cleaning

The remaining step in cleaning the data is to remove duplicates, which is applied by subsetting the names, cif and worksheet, assuming there are no two individuals with the same name in the same period acting on behalf of our clients. The columns we need for the final dataset are selected, the cost, percentage, cif and year. With the cost DataFrame obtained, it can be connected to an Excel spreadsheet which includes the client data.

Other cost brackets

The subcontractor databases is completed with key differences to the staff costs. The labels used to define the location of the data required for this data set is 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 in 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.

# List other potential data sets you could use

I could also use the report, the justification of the tax relief that our clients are due. I could use a bag of words model on this, although it probably won’t yield too great of results considering that the industry has not a significant correlation with the data. I wouldn’t be able to use this per industry or per type of project at present.

# Explain your initial findings

Costs brackets

The most important variables were to explore the expenditure itself which construct the total costs. There is significant variation in the size of each cost type; staff costs are significantly greater than software expenditure. 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 each cost type where they are present, therefore this should not disrupt the data 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 also a lot of variance within the data, with mostly small figures and large ‘outliers’, the data is exponentially distributed. It would seem that the majority of our clients would be completing ‘side projects’ and a few larger companies are generating a lot of 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 was 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. This leads to discussion of how well the characteristics of the expenditure could be used to predict total expenditure.

How much was apportioned from total staff costs for example, and how well does this translate to total expenditure? Using the median 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 on its own it does 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. The other three cost brackets have moderate correlation to the total costs in this method. The count certainly should be an useful contribution to the prediction.

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. 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. Postcodes was another feature I had 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 in how few records there are per post code. Of the top 15 most frequent post codes, there is only 643 records. There is indeed some regions with very high values but the data was not statistically significant. The final variable which I had high expectations for was the phase, the number of times a client has worked with us. 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 a significant difference, and there is nowhere near as much data 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, but also look deeper into accounting data to search for predictors.

# Deeper Analysis – Machine learning

Regression models

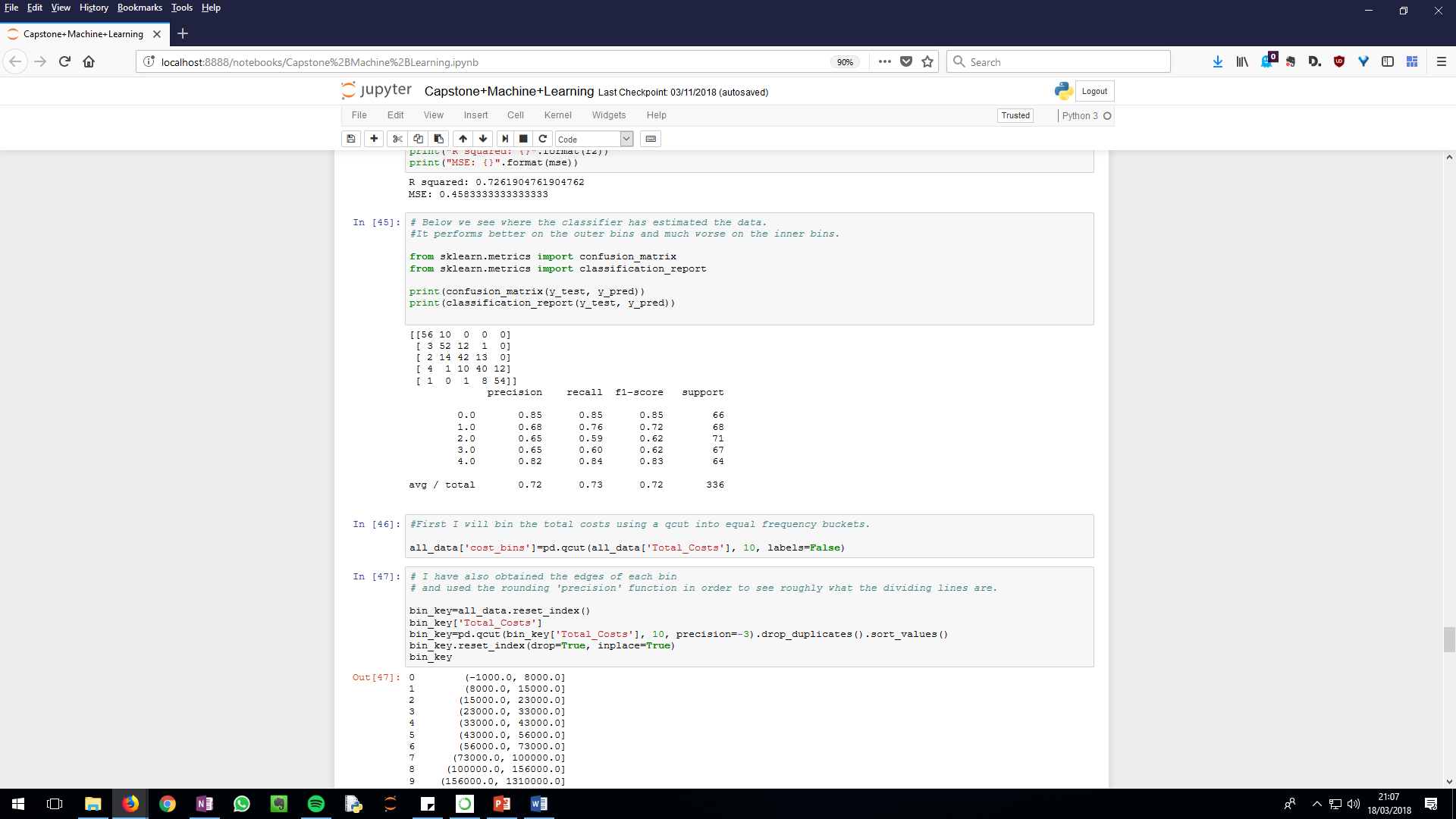
In the most ideal situation, we would be able to predict exactly how much the R&D value would be based on the client’s data. With this goal in mind, I first began to analyse the data according to regression models. The most important feature according to the inferential statistics analysis was the staff 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 was quite valuable for a single feature, returns a statistically significant P value of 0.00. There is evidence that this feature would be a powerful predictor as has been shown previously, however I would aim to construct a model without this feature. Without this feature we can use the model before a prospective client needs to thoroughly consider their projects. I therefore first analysed a linear regression model and cross validate the results. The results were very poor. We already know that using all of the features, the data is not linear or normally distributed. This is especially shown in the performance of the linear regression algorithm with an average 5 fold score of -329531. A random forest regressor is a preferable choice. Selecting 20 estimators 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. The predictions and 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 the features which had the highest P values, to remove that which is the least statistically significant. I removed the cif employees, received date, completed date 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 an improvement upon the previous test. The high p value features have bee removed but the r squared score is maintained at 0.839 and a overall 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. Scoring the regression model returns a score value of 0.71. 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. 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. 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 values would eschew the predictive power of the model. 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. In this way we can create more precise predictions to the lower and more numerous total cost values, and it is in this area the predictions would create more power – if it is over or under £20,000, this would be important for a prospective client to know whether it is worth their time. If for example the prediction is over £100,000 then a client would know that there will be a sizeable return on investment.

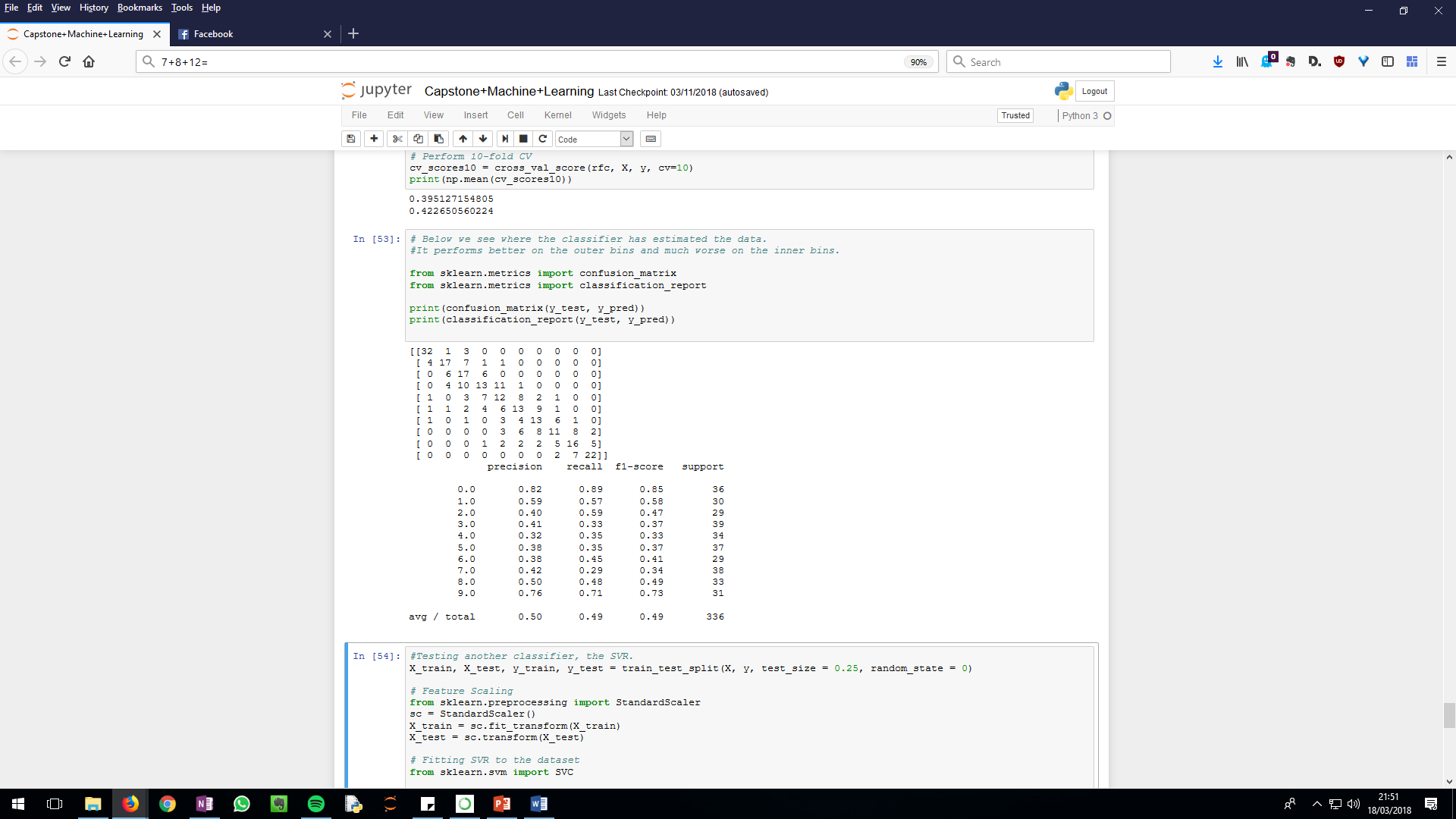
The first classification model is based upon five bins, therefore the predictions could effectively be ranked from 1 to 5 on their values. I have printed to the nearest thousand what each bin means so that the bins can be easily visualized, for example 0 is 0-£15,000, and 5 is 100,000.00 to £1,310,000.00. 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 a classification system than a regressor one.

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, the more difficult it is to predict.

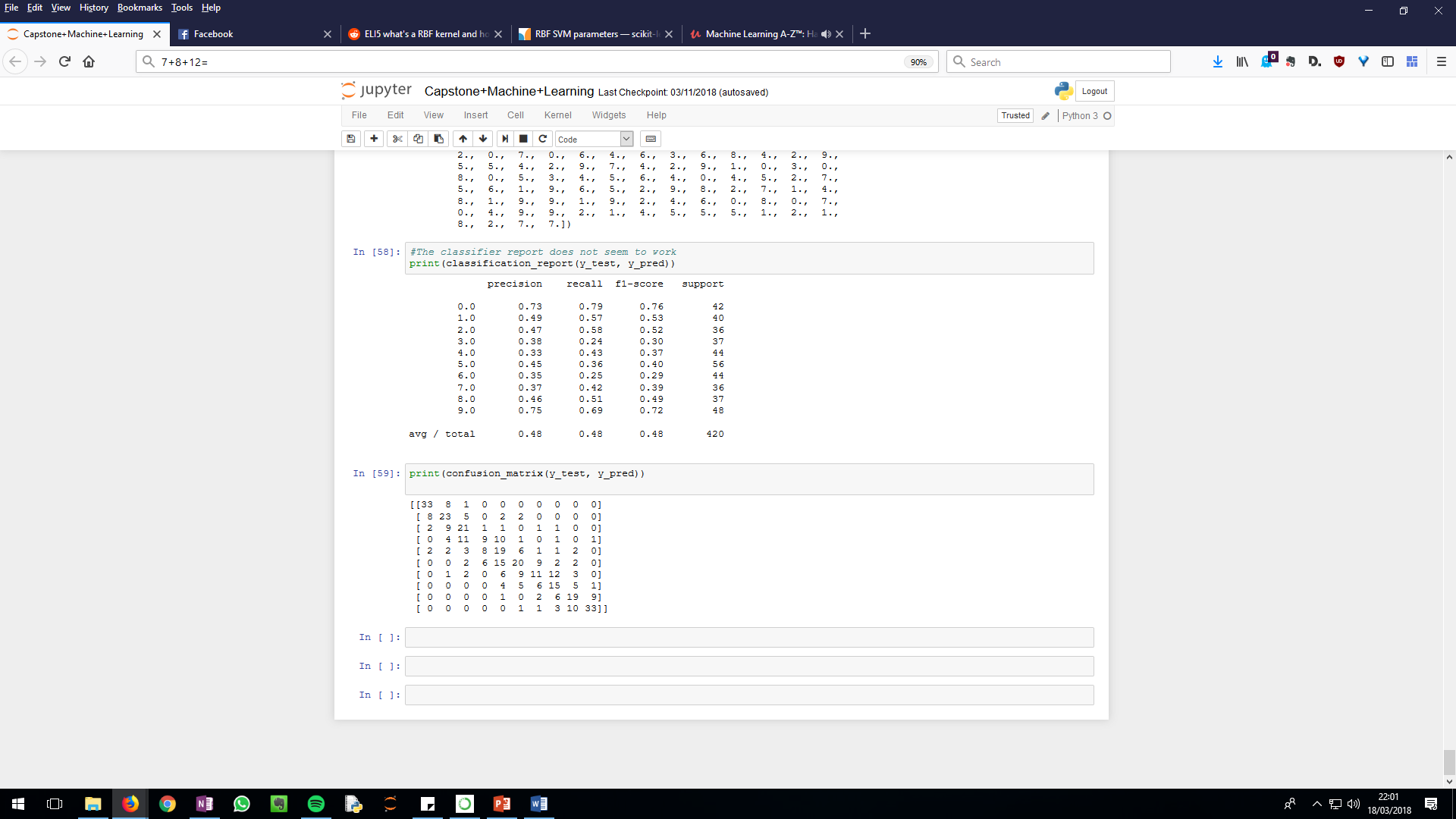


Running the classification report function shows how well the classification model has performed, and puts a figure to the description based upon the visualization above. The precision, recall and f1 scores are worse towards 2.0, but are highest at 0.0.

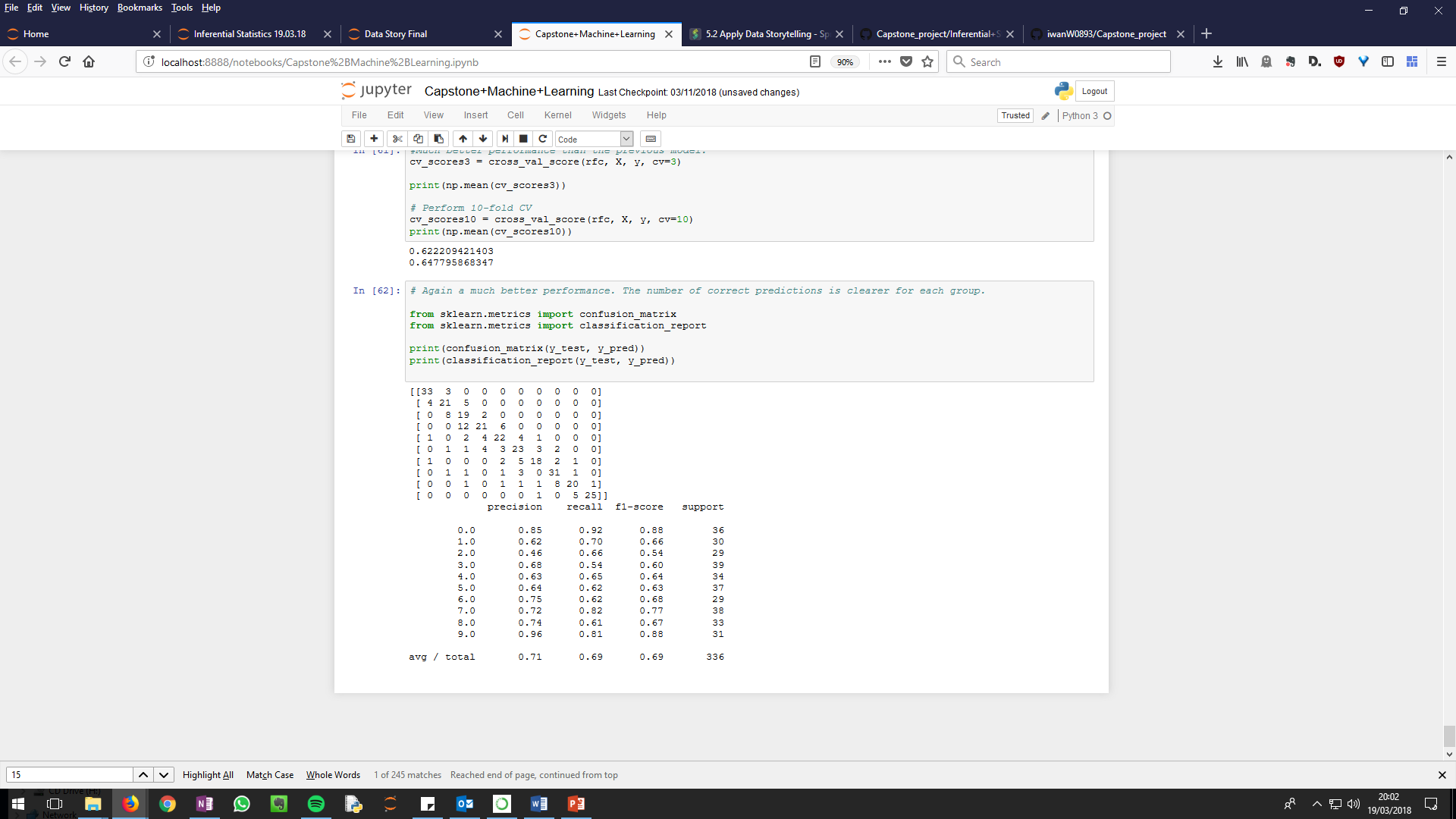
But perhaps it would be more manageable in a practical setting if there were more classification groups. I doubled the classification groups to 10. This returns a much lower R squared value of 0.49 and a higher MSE of 1.5. The cross validation score too is low, being 0.39 and 0.42 respectively with 3 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.



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. There is a lot more diversity in which class was selected for the data, which is a real weakness if we were to estimate the correct class by a range of bins. The random forest classifier is the preferred method.



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. If we were to introduce the staff costs, the strongest predictor, and pose that the client was to put some thought into how much time their staff members spent on R&D and without factoring other cost brackets. This next 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 previous predictions utilized have been rough estimates, I would recommend to test 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. This calculator is best used with the client and also utilizes some quite basic information, therefore has not been taken into consideration in this project.

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 added to it. We particularly need to pay attention to the ‘CIF’ features, which were only 344 rows out of 1678. The CIF turnover was a useful feature although not the strongest. It should be ensured that it is added to the data set with future clients. More features could also help to increase the accuracy of the data therefore this should be kept in mind. A notable feature I would further investigate is the industry categorical data. The industries that the clients have been sorted to have 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 and 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. Restructuring these industries should take greater thought than manipulating the industries for the prediction models, therefore I would leave this for the client to decide. Another method could be to break up the industries by the projects that they complete. This might be a material based project or a software based project 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, it might not be as lucrative. The models have the strength to predict R&D expenditure, however I believe that the recommendations above would greatly improve its performance.