

**Data Science for Business Final Exam**

Carla Kim Gaieski

Creative Technology Management, Yonsei University

CTM2014.01-00: Data Science for Business

Professor Keeheon Lee

June 23, 2022

<b>The Problem Set</b>	<b>2</b>
<b>Cases</b>	<b>3</b>
Case 1	4
Case 2	8
Case 3	10
Case 4	12
<b>Conclusion</b>	<b>12</b>
<b>References</b>	<b>12</b>

## The Problem Set

The problem set provided three different cases for the application of TF-IDF, SVD, Logistic Regression, and AUC-ROC curve. In each scenario the target dataframe, to be used together with the matrix, converted into a dataframe, from the SVD, changed. This way, it would be possible to generate three different logistic regression models to give a better accuracy of class probability, as a function. After the implementation of the regression, it is possible to calculate the accuracy of it numerically and by using a curve. Instead of using multiple confusion matrices for each threshold that we considered important, a ROC (Receiver Operator Characteristic) curve can summarize the information of the matrices.

In the curve, the Y-axis indicates the True Positive Rate (which is equivalent to the sensitivity;  $\text{true positives} / (\text{true positives} + \text{false negatives})$ ), that indicates the proportion of the samples that were correctly classified. And the X-axis indicated the False Positive Rate (which is equivalent to  $1 - \text{Specificity}$ ;  $\text{false positives} / (\text{false positives} + \text{true negatives})$ ). The dotted line in the ROC graphs indicates where the True Positive Rate is equal to the False Positive Rate, which means that the proportion of each correctly and incorrectly classified sample are the same. The further a curve can get from the diagonal dotted line, the better is it, as it indicates a decrease in the classification of false positives, being a better threshold. So, whenever the graph only touches the Y-axis, it results in no false positives, and those points are to determine the optimal threshold. With the completed ROC curves, it is possible to then use AUC (Area Under the Curve) to compare each ROC curve with one another and choose which categorization is better. The curve with the greater area is the better curve to consider for threshold and correct classification. The closer the area is to 1, the better, and for it to be considered a good classifier, it should be in between 0.5 and 1, as it shows that the model can classify the positives and negatives well. If equal or lower than 0.5 the model is not good, and not able to distinguish between the positives and negatives, meaning that it cannot predict well.

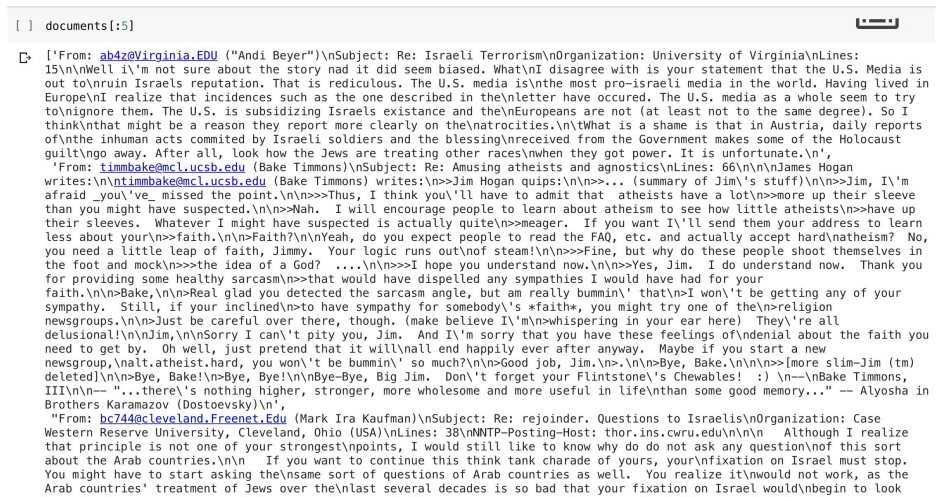
With logistic regression it is also possible to get the coefficient of the regression, which can be used to indicate the feature importance. The feature importance provides the information on the strongest features of the dataset more relevant to get to the target.

Link to the colab:

[https://colab.research.google.com/drive/18hp\\_H39mAZZTEJBJ5Q-DfzX2ukb7JdXJ?usp=sharing](https://colab.research.google.com/drive/18hp_H39mAZZTEJBJ5Q-DfzX2ukb7JdXJ?usp=sharing)

## Cases

The procedure for all cases were the same. The data frame containing documents and terms was cleaned, meaning that the data frame had any information or written aspect of it (replace special character with blank space, split into words, that if greater than three, it will be joined, and lastly all letters were lower cased) changed for easier manipulation of it.



```
[ ] documents[:5]

[ ] From: ab4z@virginia.edu ("Andi Beyer")\nSubject: Re: Israeli Terrorism\nOrganization: University of Virginia\nLines: 15\nWell I'm not sure about the story nad it did seem biased. What\nI disagree with is your statement that the U.S. Media is out to\nruin Israels reputation. That is ridiculous. The U.S. media is\nthe most pro-israeli media in the world. Having lived in Europe\nI realize that incidences such as the one described in the\nletter have occurred. The U.S. media as a whole seem to try to\nignore them. The U.S. is subsidizing Israels existance and the\nEuropeans are not (at least not to the same degree). So I think\nthat might be a reason they report more clearly on the\natrocities.\nWhat is a shame is that in Austria, daily reports of\nthe inhuman acts committed by Israeli soldiers and the blessing\nreceived from the Government makes some of the Holocaust guilt\ngo away. After all, look how the Jews are treating other races\nwhen they got power. It is unfortunate.\nFrom: timmbake@mcl.ucsb.edu (Bake Timmons)\nSubject: Re: Amusing atheists and agnostics\nLines: 66\nJames Hogan writes:\n>Bake Timmons writes:\n>>Jim Hogan quips:\n>>>... (summary of Jim's stuff)\n>>>Jim, I'm afraid you've missed the point.\n>>>Thus, I think you'll have to admit that atheists have a lot\n>>>more up their sleeve than you might have suspected.\n>>>Mah. I will encourage people to learn about atheism to see how little atheists\n>>>have up their sleeves. Whatever I might have suspected is actually quite\n>>>meager. If you want I'll send them your address to learn less about your\n>>>faith.\n>>>Faith? Yeah, do you expect people to read the FAQ, etc. and actually accept hard\n>>>atheism? No, you need a little leap of faith, Jimmy. Your logic runs out\n>>>steam!\n>>>Fine, but why do these people shoot themselves in the foot and mock\n>>>the idea of a God? ....\n>>>I hope you understand now.\n>>>Yes, Jim. I do understand now. Thank you for providing some healthy sarcasm\n>>>that would have dispelled any sympathies I would have had for your faith.\n>>>Real glad you detected the sarcasm angle, but am really bummin' that\n>>>I won't be getting any of your sympathy. Still, if your inclined\n>>>to have sympathy for somebody's faith, you might try one of the\n>>>religion newsgroups.\n>>>Just be careful over there, though. (make believe I'm\n>>>whispering in your ear here) They're all delusional!\n>>>Jim,\n>>>Sorry I can't pity you, Jim. And I'm sorry that you have these feelings of\n>>>denial about the faith you need to get by. Oh well, just pretend that it will\n>>>end happily ever after anyway. Maybe if you start a new newsgroup,\n>>>alt.atheist.hard, you won't be bummin' so much?\n>>>Good job, Jim.\n>>>Bye, Bake.\n>>>[more slim-Jim (tm) deleted]\n>>>Bye, Bake!\n>>>Bye! Bye-Bye, Big Jim. Don't forget your Flintstone's Chewables! :) \n>>>Bake Timmons, III\n>>>...there's nothing higher, stronger, more wholesome and more useful in life\n>>>than some good memory..." -- Alyosha in Brothers Karamazov (Dostoevsky)\nFrom: bc744@cleveland.freenet.edu (Mark Ira Kaufman)\nSubject: Re: rejoinder. Questions to Israel's\nOrganization: Case Western Reserve University, Cleveland, Ohio (USA)\nLines: 38\nNNTP-Posting-Host: thor.ins.cwru.edu\nAlthough I realize that principle is not one of your strongest\npoints, I would still like to know why do do not ask any question\nof this sort about the Arab countries.\nIf you want to continue this think tank charade of yours, your\nfixation on Israel must stop. You might have to start asking the\nsame sort of questions as well. You realize it\nwould not work, as the Arab countries' treatment of Jews over the\nlast several decades is so bad that your fixation on Israel\nwould\nbegin to look
```

Figure 1. The documents used in the problem

	document	clean_doc
0	From: ab4z@virginia.edu ("Andi Beyer")\nSubjec...	from virginia andi beyer subject israeli terro...
1	From: timmbake@mcl.ucsb.edu (Bake Timmons)\nSu...	from timmbake ucsb bake timmons subject amusin...
2	From: bc744@cleveland.freenet.edu (Mark Ira Ka...	from cleveland freenet mark kaufman subject re...
3	From: ray@ole.cdac.com (Ray Berry)\nSubject: C...	from cdac berry subject clipper business usual...
4	From: kkeller@mail.sas.upenn.edu (Keith Keller...	from kkeller mail upenn keith keller subject p...

Figure 2. The cleaned data frame of the documents

Next, the data frame was tokenized - splitted into a series of words, grouped into a list so they could be vectorized by the TF-IDF, to measure the importance of a word, by highlighting

the words that are frequent within one specific document. Only after the TF-IDF is it possible to do the SVD (Singular Value Decomposition) to use the U matrix given in the result, that contains 1000 rows and 20 columns. This is the matrix, turned into a new data framed, used as the X in the regression models for all three cases.

	0	1	2	3	4	5	6	7	8	
0	0.122623	-0.099399	-0.027604	0.018124	0.066939	0.010938	-0.030702	-0.005237	0.159802	-0.0508
1	0.204797	-0.133624	-0.090127	-0.085361	0.047217	0.010852	0.027775	-0.046539	-0.111683	0.0178
2	0.240022	-0.134487	-0.045289	0.074205	0.182686	0.058706	-0.025885	-0.006377	0.262795	0.0143
3	0.165554	-0.030227	-0.092470	0.114398	-0.171096	-0.194415	-0.093528	0.040157	0.015350	0.0485
4	0.213076	0.032175	0.095634	0.039444	0.105356	0.016778	-0.044837	0.036805	-0.143326	0.0190
...	...	...	...	...	...	...	...	...	...	...
11309	0.073887	-0.008979	0.049855	0.066923	0.091830	0.024522	-0.040726	0.036097	0.039984	0.0385
11310	0.145940	-0.005989	0.032423	0.050419	0.061806	0.016070	-0.016299	0.003586	-0.056062	0.0460
11311	0.172356	-0.066974	0.072968	-0.075805	-0.050570	-0.001756	0.006255	-0.001912	0.017824	-0.0011
11312	0.197381	0.105903	0.083140	0.111590	0.098364	0.003600	-0.034019	0.040118	-0.095339	0.0343
11313	0.240448	-0.066653	0.075166	-0.076579	-0.081472	-0.009830	-0.046703	-0.020108	0.011077	0.0240

11314 rows x 20 columns

Figure 3. Transformed data frame to be used in the logistic regressions

### Case 1

For the first case, the target data frame was already given, and it is this target that was used in the logistic regression mode as the Y component, as it is a single column with 20 rows.

target	
0	17
1	0
2	17
3	11
4	10

Figure 4. The first target data frame

Once the logistic regression is run between the X and Y it was possible to get the accuracy coefficient of the model, which was of approximately 56%. The accuracy of 56% is not the greatest nor the desired result to accomplish when performing a regression analysis. However, given the time for producing the model, it is not a bad result, which leaves room for improvement for the model, to elevate its accuracy.

When making the ROC curve, it is possible to plot all of the thresholds together in the same graph, which makes it possible to compare one class(variable) to all the others. This curve can already indicate that the model does not perfectly distinguish between positives and negatives, but there should be at least one class with a high AUC.

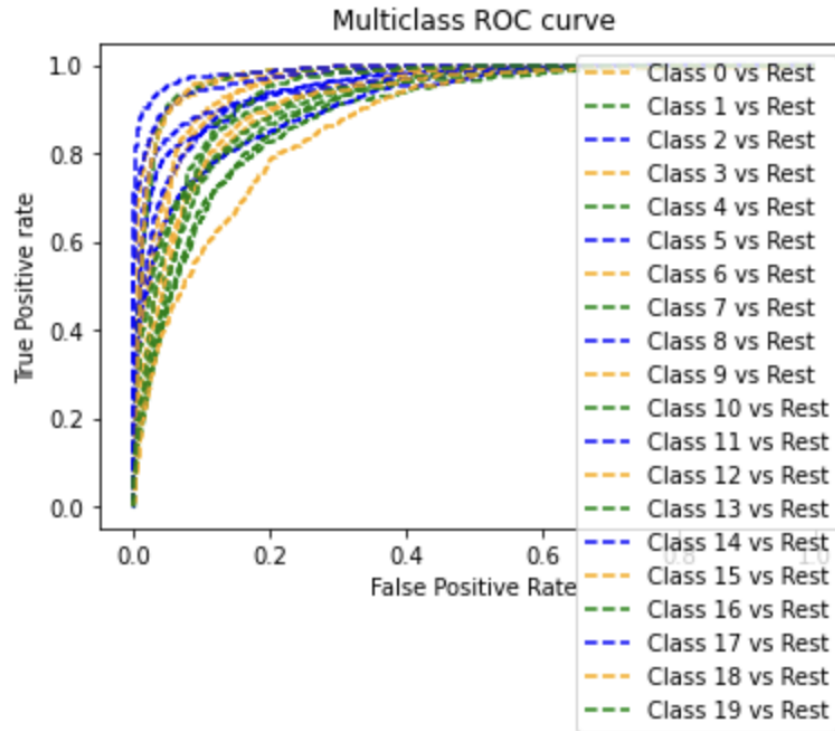


Figure 5. Multiclass ROC Curve

Since it is a multiclass problem, the multiclass ROC curve makes it harder to evaluate each feature to its fullest. So, it is necessary to plot individual ROC curves for each class that the dataframe has. When plotting all the different class ROC-AUC curves, there are many curves with the area of 0.80, but there is one class which has an AUC of 0.99, being the highest of them all. The feature with the highest area is the feature 17. The area closest to 1 is the model that can perfectly classify the positives and negatives with precision, therefore, the feature 17 is the most relevant in this case, with a high importance, due to its almost perfect classification in this model.

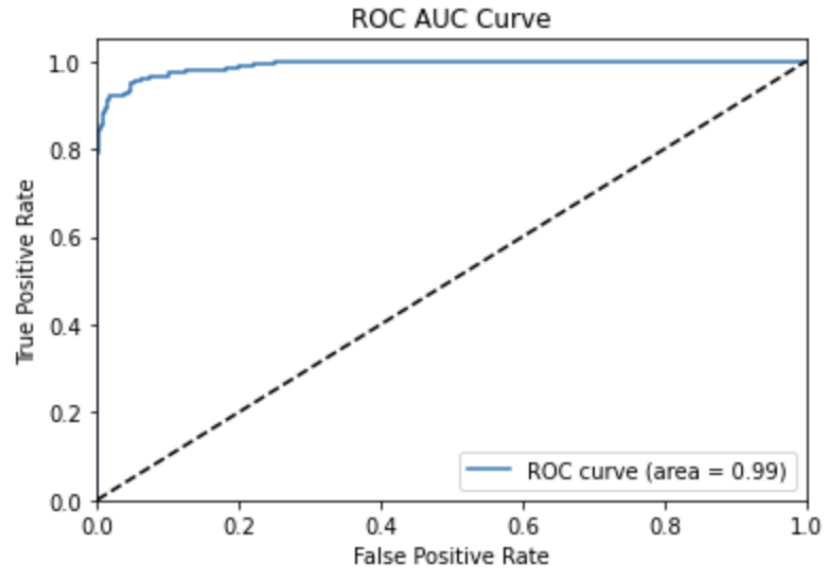


Figure 6. ROC curve with the highest AUC of 0.99 (feature 17)

When analyzing the plot for the feature importance, no clear pattern can be extracted from it. For further analysis and improvement of the model it would be necessary to spend more time, and try out other methods of evaluation of the regression.

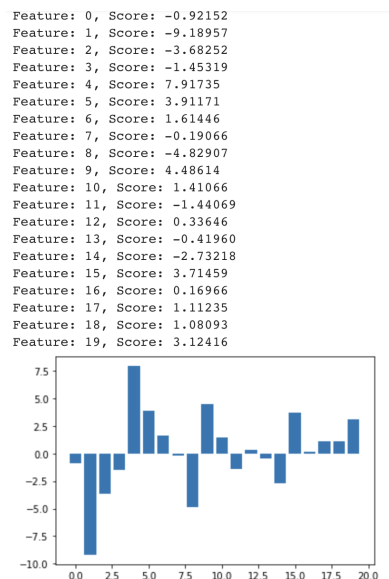


Figure 7. Feature importance



## Case 2

For the second case, it was necessary to create a new target dataframe, to be used as the Y in the logistic regression. The new target dataframe was based on the sender's e-mail, if it was academic or not, meaning if the sender's e-mail finished with “.edu” or not. This made it possible to create a binary dataframe where 1 indicated academic and 0 non academic.

academic	
0	1
1	1
2	1
3	0
4	1

Figure 8. New target df

After running the same process of logistic regression, it was possible to achieve a higher coefficient for this model, of approximately 66%. Compared to the first case, this shows a significant increase in the accuracy of the model, with an improvement of around 10%. This also indicates that the new target dataframe is better to be compared with the document dataframe, as there is more accurate prediction with this feature.

The AUC in the ROC curve in this case is of 0.72, which is not greater than the one from the specific feature in the first case, but it shows good classification of the positive class in the dataset, from the regression model. This is because the AUC value is in between 0.5 and 1. With the feature importance, it was also possible to see what correlates more to the target dataframe, given their individual scores.

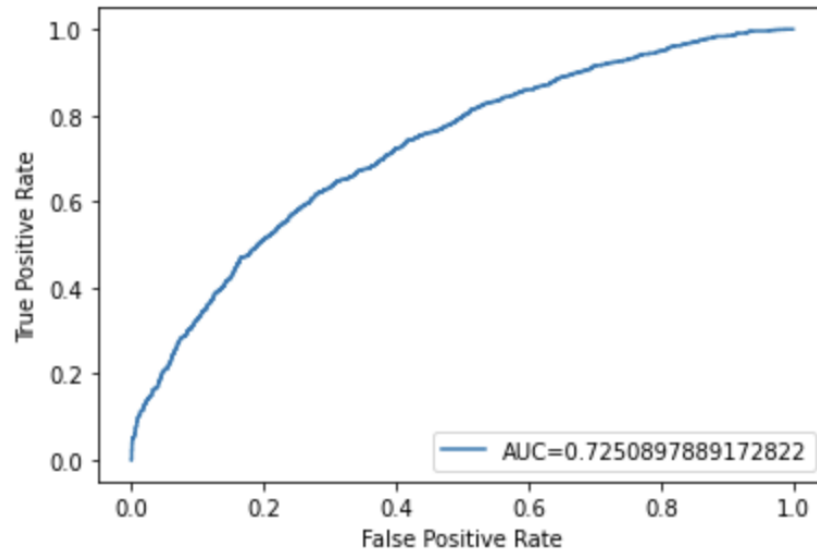


Figure 9. Regression case 2 ROC- AUC curve

The important features are positive and negative, as they indicate 1 and 0 features, respectively. This comes to show that there are features with a higher score, but still makes it hard to extract a pattern from. For further analysis and improvement of the model it would be necessary to spend more time, and try out of other methods of evaluation of the regression.

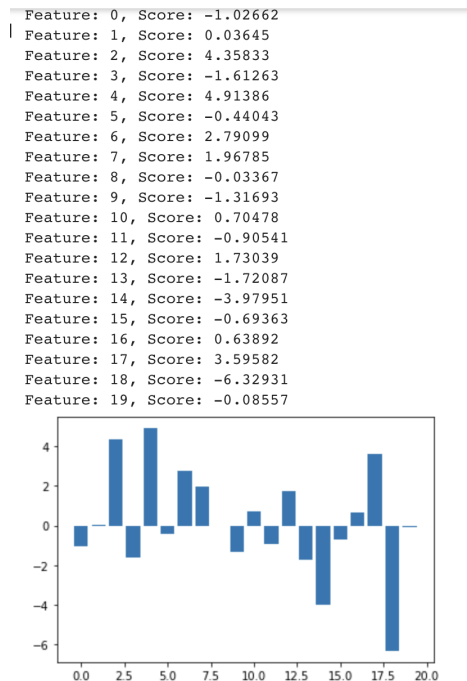


Figure 10. Feature Importance for case 2

### Case 3

The third case also required the creation of a new target dataframe, that was determined by the receiver's email. This, however, generated an issue in the length of the data frames, as with further inspection of the documents dataframe, some e-mails were missing the “To” part, which made them a NaN. A NaN, when being dropped caused the issue of the entire row being deleted, which created the imbalance when running the regression. So, once again, it was transformed into a binary dataframe, eliminating the NaN rows. So, the documents dataframe had to be modified to also eliminate the NaN rows, so when combining both dataframes the length would match.

academic	
0	1
1	0
2	1
3	1
4	1
...	...
9327	1
9328	0
9329	1
9330	0
9331	1

Figure 11. New target df

Running the logistic regression led to the achievement of a coefficient of 62%, which for a regression that started first at 20%, it is a great improvement. This comes to show that the logistic regression model of the receiver's e-mail as the target also shows a somewhat good classification.

The ROC curve provided an AUC of 0.67, which is also a good number. It is between 0.5 and 1, which also means that the classifier can distinguish between the positive and negative points.

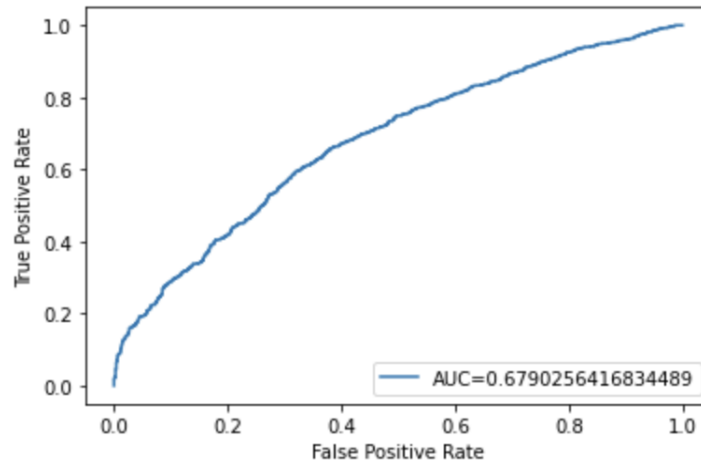


Figure 12. ROC-AUC curve for case 3

The feature importance also shows no pattern, but the scores are much lower than the previous cases. For further analysis and improvement of the model it would be necessary to spend more time, and try out of other methods of evaluation of the regression.

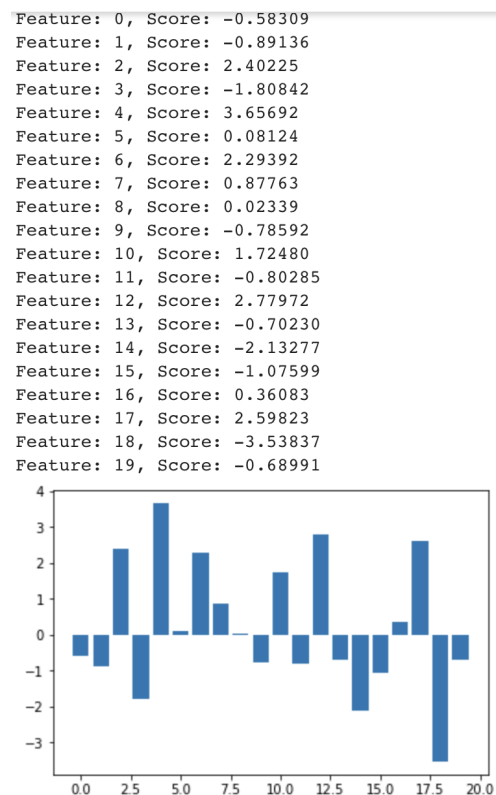
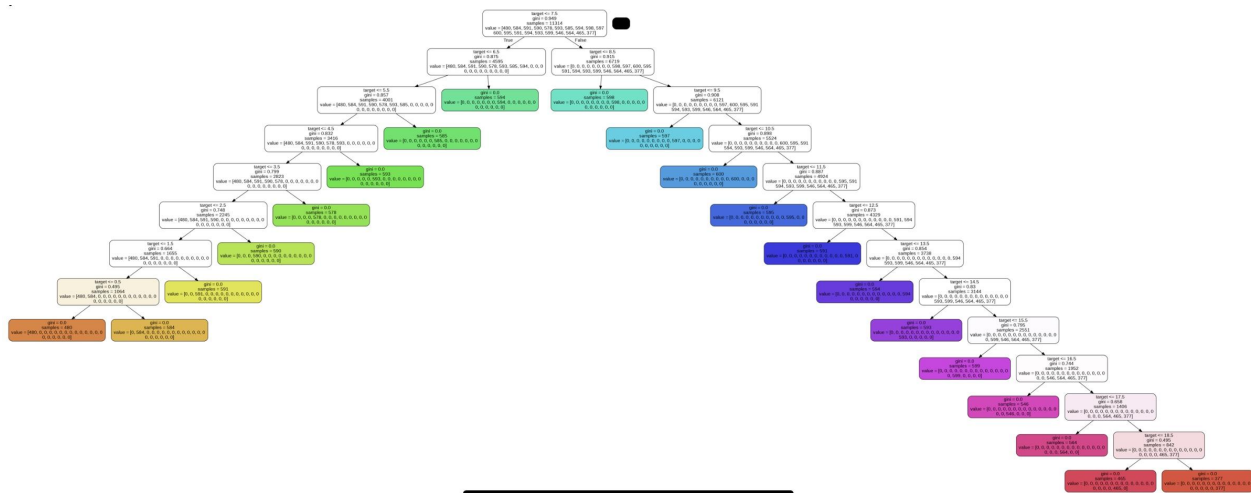


Figure 13. Feature importance for case 3

## Case 4

Case 4 required more time for development and analysis, in order to create decision trees and individual explanation. However, it was only possible to plot a decision tree for the first case, with the dataframes that were already given in the problem set. It shows many branches and nodes. All of the nodes carry the Gini index of 0, which cancels any uncertainty that could have on the classification made with the tree, as it means no impurity. The targets that keep on diving have Gini's that are not 0.0, and until they meet certain conditions, it will keep branching downwards.

## Case 1



**References**

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (1st ed.) [E-book]. O'Reilly Media.