Craig Phifer
9/19/2018

Enron Submission Free-Response Questions

A critical part of machine learning is making sense of your analysis process and communicating it to others. The questions below will help us understand your decision-making process and allow us to give feedback on your project. Please answer each question; your answers should be about 1-2 paragraphs per question. If you find yourself writing much more than that, take a step back and see if you can simplify your response!

When your evaluator looks at your responses, he or she will use a specific list of rubric items to assess your answers. Here is the link to that rubric: [Link] Each question has one or more specific rubric items associated with it, so before you submit an answer, take a look at that part of the rubric. If your response does not meet expectations for all rubric points, you will be asked to revise and resubmit your project. Make sure that your responses are detailed enough that the evaluator will be able to understand the steps you took and your thought processes as you went through the data analysis.

Once you've submitted your responses, your coach will take a look and may ask a few more focused follow-up questions on one or more of your answers.

We can't wait to see what you've put together for this project!

1. Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: "data exploration", "outlier investigation"]

This investigation sought out to examine the dataset in order to uncover financial discrepancies that could lead us to identify Enron employees, looking at potential Persons of Interest and other employees, who may have committed fraud at the time of the company's dubious reporting practices. Investigating this imperfect and real-world dataset by cleaning it of unessential features and outliers to avoid overfitting of our refined data, then making use of machine learning algorithms to develop and tune classifiers of POIs and applying it to the larger individual population for further identification opportunity. Machine learning helps us to analyze and explore the data much faster than we could do independently for each of the many features, and in most cases more accurately.

In the dataset we identified 22 total features and 146 individuals, 18 individuals were labelled as POIs. The POIs were used as test labels in our algorithms which were trained to be compared with the remaining individuals to identify for feature similarities. The features were refined by intuition based on maximizing the prediction of POIs. Comparing bonus and total stock value features allowed us to visualize the data points in a scatter plot in which identified three major outliers including “Total”, “The Travel Agency in the Park”, and “Eugene Lockhart”. Once identified as not being individuals working at Enron or who had no data reported, we simply removed them from the data dictionary.

2. What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: “create new features”, “intelligently select features”, “properly scale features”]

In determining an appropriate features list, knowledge and intuition about the Enron scandal and how executives gamed the system by to jack up stock prices while taking advantage stock options was initially used. Based on this understanding of Enron’s actions and the three types of features included in the dataset (Persons of Interest, Financial, and Email) it was fairly straightforward to that we would rely more on the POI and Financial features initially. We also selectively chose to limit features based on the amount of missing values.

Using intuition on the details of the Enron fallout, a ratio between stock and salary was deployed to potentially identify a new feature. Ultimately this feature did not work out so well. SelectKBest was also used to determine which of our features was best used and this new feature returned a low score of 0.119. With the averting of overfitting in mind, four features with a KBest value score below 5 were removed. Our final features list includes:

- poi
- salary
- total_payments
- bonus
- total_stock_value
- exercised_stock_options
- from_poi_to_this_person
- shared_receipt_with_poi

In using KNearestNeighbors as the algorithm of choice for our classifier feature scaling was unavoidable. It was incorporated into Pipeline via MinMaxScaler. This estimator scales and translates each feature individually such that it is in the given range on the training set, normalizing the data. This scaling provides a very reliable number as to what we can expect in the output. Without this step inconsistencies in the values of the scales for both axes will result in poor performance.

3. What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]

In Task 4 we employed the use of four classifiers, all of which were introduced in the Machine Learning Lesson sets. They include Naive Bayes (GaussianNB), Support Vector Machines (SVC), K-Nearest Neighbors (KNeighborsClassifier), and Decision Trees (tree). Ultimately K-Nearest Neighbors was the algorithm of choice for further parameter tuning. The accuracy scores of all the algorithms explored are below

```
NB Accuracy: 0.859154929577
SVM Accuracy: 0.845070422535
Accuracy_KNC: 0.845070422535
DT Accuracy: 0.788732394366
```

4. What does it mean to tune the parameters of an algorithm, and what can happen if you don’t do this well? How did you tune the parameters of your particular algorithm? What parameters did you tune? (Some algorithms do not have parameters that you need to tune -- if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric items: “discuss parameter tuning”, “tune the algorithm”]

Part of the artistry of Machine Learning, as gleaned from the lesson references, is to tweak the algorithm in such a way that it produces the appropriate balance between bias and variance. It involves tuning parameters (features and labels) so that you are not overfitting your data. A high-bias variance machine learning algorithm is one that practically ignores the data. It has almost no capacity to learn anything. In the other direction, making the algorithm extremely perceptive to data makes it high variance. The issue with this is that it will react poorly to situations it has never encountered before because it lacks the right bias to challenge new data.

We tuned the parameters of the K-Nearest Neighbors algorithm with GridSearchCV. It helped us identify the best fit for the given n_neighbor, and weights parameters ('n_neighbors': 6, 'weights': 'uniform').

5. What is validation, and what's a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric items: "discuss validation", "validation strategy"]

Validation occurs when we test our algorithm, making sure that our results are actually trustworthy. This process shows up in the form of the test/train split, k-fold, and visualizations of our data. Throughout the project we include visualizations including scatterplots to understand this representation of our data at various points.

Perhaps the most significant use of validation takes place in Task 5 where we tune the parameters of our algorithm to improve precision and recall. Using such a small dataset in this project can potentially make for a volatile model. Accounting for this understanding StratifiedShuffleSplit cross validator was used to validate our analysis, which preserved the percentage of samples for each class.

This was the most difficult aspect of this project and likewise the most crucial. In the end, with the adjustments made we even able to improve slightly the accuracy of the algorithm. With the results provided from the tester.py file we can feel confident about our validation process.

Accuracy: 0.86267
Precision: 0.47642
Recall: 0.30300

6. Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm's performance. [relevant rubric item: "usage of evaluation metrics"]

The precision metric returned a score of 0.47642, which can be rounded to 48%. This metric is the fraction of relevant instances among the instances retrieved, and can be explained as the persons who were identified as POIs actually being POIs. About 52% of persons the model had identified were not relevant. There are potentially many reasons for this result, perhaps one result is that identifying POIs is not a guaranteed process. Even the best detectives are still human and make mistakes. This affects the original labelling of POIs.

The recall metric returned a score of 0.30300, which can be rounded to about 30%. Recall is also known as sensitivity, measured as the fraction of relevant instances that have been retrieved over the total amount of relevant instances. We only had 30% of relevant POIs selected in our model.