Data Science Portfolio Project Review

Aiden V. Johnson
December 30, 2018
Project #1 Review

This is the model review for Reshad Reza project #1. In general the project follows the standard
Data Science Method, which contains the following steps:
Steps highlighted in red need work, those in green will vary depending on the project model implementation needs.
1. Problem Identification
2. Data Collection, Organization, and Definitions
3. Exploratory Data Analysis
4. Pre-processing and Training Model Data Development
5. Fit Models with Training Data Set
6. Review Model Outcomes—Iterate over additional models as needed.
7. Identify the Final Model
8. Apply the Model to the Complete Data Set
9. Review the Results—Share your findings
10. Finalize Code and Documentation

Technical Comments

Please include at least one sentence in the Jupyter notebook about the goal of the modeling project at the beginning.

It is a good practice to review the NA's and data types, and general description of the data such as the `pandas.describe()` function as you did. Please, handle any data type transformations at this point as well to stream line the EDA by using what is considered the * Model Development Dataset*. Performing variable review on data that is not in the format you will use in model development can result in spurious findings informing your initial model development work.

Following the Model Development Dataset process create dummy variables for all non-numeric columns such as *salary* before performing EDA steps such as the correlation heat map. Complete this step for all categorical variables.

Also the *promotion_last_5years* variable also is non-numeric as it indicates a True/False boolean. The way it is currently incorporated in the correlation heatmap is incorrect. Double check that all True/False indicator variables are represented as such before perfomring EDA including the correlation heatmap.

Many of the other variables are hard to evaluate because there is no description about what they are, such as *satisfaction_level*, which is probably a percentage response from employee surveys based on our conversation. However, if you are not going to include a project report then you need to provide some details within the Jupyter notebook to guide the reviewer. Ideally, the project should be able to stand on it's own and be understood by anyone with a data science background without having a conversation with you before hand.

Please review the section of this article on Data Collection and Data Definitions. There are many different names for these practices, but applying some form of these is a must. Consider what steps you could take to add comments and documentation to guide the reviewer.

In the K-means clustering you failed to use an appropriate method for the number of clusters to consider, such as an elbow plot or silhouette width. You can also apply a test and check approach, a decision that k = 3 must be a logical choice. Plus, what was the purpose of the clustering, you didn't use these clusters for any further modeling work or discuss them.

Modeling work is solid as is the comparison and performance review, more on this in specific comments.

Reviewing the variable importances in detail would be very informative. Take the top five or ten variables and determine the important values for leaving the company. What is the definition of *time_spend_company* ? My guess is that people don't leave after they have been with the company for over 5 years, however, you need to get clearer on the variables before statements like that can be identified.

Specific Notes

line 3 - Why do you have a `.csv.txt` it is either a comma separated value file or text file? Pick one and rename the input accordingly.

line 10 - mean similarity should be discussed or creating uni or bi-plots of the response variable against each explanatory variable would also be interesting.
Try `seaborn.pairplot(df)`

line 12 - Why did you choose these variables to plot? What do the plots indicate?

line 13 - Is `0` or `1` indicating they left? This looks like most employees did not leave, is that correct? Why are those that stayed mostly in three groups, and there is something systematic going on with the two tight orange boxes, Last evaluation is what a date?

line 14 - why k-means on only `left==1` and `"satisfaction_level","last_evaluation"` ?

line 15 - This is just another view of line 13, not needed unless it shows something new.

line 16 - yes this is what I am referring to in line 10, more of these type plots for each explanatory variable.

line 17 - Remove this - line 18 is the same information and much easier to review. Make the labels larger or plot itself larger.

line 19 - Why not a regular histogram here? Also how does hours relate to the other explanatory variables? Can the number of hours worked be explained by specific department or salary? I am guessing those three are highly correlated.

line 20 - Good, do this earlier as mentioned

line 23 - Nice

line 26 - This is a repeat of line 23 and huge barplot is unnecessary.

lines 27-30 - Resample is good as is Train, Test split.

lines 31 - 38 - Nice model development work.

line 39 Neat, but how useful is this? The Employer is probably not worried about False Positives and with those AUC scores you don't need to show this. Obviously, you will choose either the Random Forest or the Gradient Boosting model based on performance.

lines 40 - 42 - Pick the final model and print the CM for that model.

line 43 - Remove this

line 44 - This is where the take home message is. Reformat these to relative importances between 0 - 1. Also the labels are very hard to read, make them larger and/or brighter.

Project Report Comments

There is currently no project report to provide comments on.