Predictive Analytics

OMIS 115

Project

Goal: The goal of this project is to apply whatever you have learned in this course on a real-world problem.

Deliverables:

- Code (Jupyter)
- Data used (If more than 100M, use google drive and share the file)
- Power point presentation

Problem:

You need to find / select a problem that requires prediction of a variable via a classification or regression model. Try to clearly define the problem. Specify all the assumptions made. Try to an find interesting problem!

Data:

Should be not too small and not too large.

(For example, it can have between 10 and 100 attributes and between 1000 and 10,000 samples/instances).

You can find interesting problems and datasets online. For example, the following sources have interesting data:

- UCI Machine Learning Repository (contains for than 470 datasets)
- Kaggle (more than 1600 datasets)
- Data.gov

Model Development:

1) **Investigate your Data:** Analyze your data. Find the relationship between your features. Use descriptive statistics to learn about your data. Present these statistics in your presentation. If possible, plot your data or part of your data. You need to do preprocessing if your data requires. Talk about how, if, you had to deal with missing data, outliers etc. Understanding the data is one of the most important steps that can guide you in the next steps (e.g. model selection, feature selection, evaluation metrics etc). See if you need to do normalization.
2) **Model Selection:** You should try different models and select the best. You need to justify why you chose whatever model you chose at the end.
3) **Feature Selection/Engineering:** If you must do feature selection/engineering, present and justify it.
4) **Parameter Tuning:** If your model has parameters, you need to tune them. Find the best set of values for your parameters.
5) **Training and Testing:** When splitting your data for training and test, split it into training, validation and leave a data set for the final test and evaluation of your model. Using cross validation is recommended.
6) You need to take all the possible measures to avoid data leakage.
7) **Evaluation metrics:** You need to choose the right evaluation metrics to evaluate your model. You should be able to justify your evaluation metric choices.

   Also, compare your final model with a ***Dummy Predictive Models***. Of course, your model should outperform the dummy model.
8) **Predictive Errors:** Analyze your prediction errors. Talk about variance and bias.
9) **Results:** Finding interesting results is a plus. Also, visualization is recommended and a plus for the whole process, wherever possible.

In your presentation, talk about underfitting or overfitting problems that you encountered and how you overcame them.

You need to talk about your code and run all or part of it in class. It should run with no error. Your code should be written in Jupyter and should be organized, with headings, sub headings, comments and descriptions.

In your presentation, at the end, talk about your experience throughout the project. What did you learn about predictive analytics and machine learning from working on this project. Talk about your experience with different models that you tried. Can you give recommendation about these models given your experience? Does your experience tell you that one class, if any, is better than the other, at least for a specific task?

Have fun and good luck!

Dr. Eghbal Rashidi