Machine Learning Project, Spring 24

Due Sunday May 5th , 11:59pm.

1. Introduction

Your project topic must include data analysis using a machine learning model in one or more of the following categories: classification, regression, Association Rule Mining, clustering and recommender system). We did not cover Association Rule Mining and Recommendation Systems in class but you are allowed to choose a project in this category, if you like. The dataset you choose for your project could be any data you are using at work, data that is involved in your everyday life, or any publicly available dataset. You must perform a complete data analysis lifecycle including data visualization and exploratory data analysis, variable selection and feature extraction, running multiple machine learning models and comparing their performances.

Please note that the scale of your project should be bigger than any of the assignments you did the class and may require you to research and learn beyond the materials covered in the class.

What you need to turn in

Here are the items that you need to submit:

1. **A project Report :** This includes an extended report of your project in Pdf and should contain figures and tables that are necessary to make the report complete. Beconcise in your writing and consult technical writing references as needed. Your project report should more or less follow the following structure:
    a. **Abstract (2 pts:** A one paragraph summarizing the problem, the method you used to do the analysis and the results of your experiment.

    b. **Problem definition and project goals (3pts):** In this section, you explain the purpose of your project and the problem you are trying to solve as well as the dataset that you used for your project. How did you obtain this dataset? What features are included in the dataset and what are you planning to do with this data?

    c. **Related Work (5 pts):** Has there been any other paper/work which addressed the same problem? If so, include a very brief description of their dataset, method and results and cite it in your report. If you are choosing a project from Kaggle, you can compare your models' performances with other notebooks posted on Kaggle.

    d. **Data Exploration and preprocessing (20 pts):** In this section, you should explain how you explored the data and the relationship between variables in your dataset. Use scatter plots, boxplots, or histograms, to visualize your data and detect possible correlations between different features and the outcome variable. How did you cleaned and preprocessed the data. Did you do any feature engineering? What variables did you use and why? Is there any missing

values in your data and how did you deal with it? Is your data imbalanced? What did you do to address the imbalance problem? You also need to explain if you have done any feature scaling, normalization, categorical feature encoding, etc.

e. **Data analysis and experimental Results (50 pt):** In this section you explain the Machine learning models you used to solve the problem, present the result of your data processing and explain how you achieved the project goal through this result. This sections hould have the following components:

You must use multiple machine learning models and evaluate them to see which one works best for your problem. At the very least the models you try must include simplermodels such as regularized linear or logistic regression, random forest and gradient boosted machines, and more complex models including neural networks. E

Explain how you tuned the hyper-parameters of each model and evaluate the results via standard evaluation measures (such as AUC, RMSE, precision, recall, etc. ((this is covered in week 11lectures)

Compare the performance of your best model against a simple benchmark (e.g., for regression, you can use a model that always predicts mean, for classification, you can use a model that always predicts the majority class) We will discuss model benchmarking in Module 13.

IF your dataset has protected attribute such as (race, gender, zipcode, age, race, religion, etc.) audit your model for fairness to ensure it has equitable outcome for different protected groups. (This will be covered in module 13)

f. **Conclusion** : In this section explain any new knowledge or interesting findings you obtained from processing your data. You can also briefly mentionany further research directions.

g. **References :** This includes the bibliography. Please list any external source (including websites, books, articles) that you used and make sure to cite them within your text.

2. **"Source code**" . You must run your source code in R notebook and turn in both your R notebook with "rmd" extension and its html file (with extension .nb.html). When you preview your notebook in r studio, it will automatically create the html file with .nb.html extension. You must submit this file as well. **It is very important that you submit both of these file extensions or your submission will not be graded or will be scored zero.**
3. **Oral Presentation (20pt) :** You need to prepare a set of slides and **record an audio presentation.**. Thefirst slide should include the title of your project and your name. The next slides should include a quick overview of the presentation. The rest of the slides should

explain your work and the results. **You must record your voice over the presentation and submit it together with the other required files.**

Looking for Sample Project and/or Datasets

Check out www.kaggle.com for some interesting challenges posted by companies and researchers. Kaggle is a platform for data analytics where companies and researchers post their data and use cases and data scientists from all over the world compete to produce the best model.

Additional Notes (Review carefully to avoid losing points):

1. Your project must include some exploratory data analysis. You must explore the relationship between your features and the outcome variable.
2. You must handle missing data if they present in your dataset. If missing data is a very small percentage of your training data, you can remove it otherwise; use an imputation method (such as the one we used in assignment 5) to replace missing values. You can find some help in these blog posts: https://blogs.oracle.com/datascience/3-methods-to-handle-missing-data and https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
3. Regardless of whether you are doing classification or regression, you must use multiple machine learning models. At the very least the machine learning models you try must include simpler models such as regularized linear or logistic regression as well as ensemble models such as random forest and gradient boosted machines, and more complex models such as neural networks. **You will lose several points if you avoid neural networks as they require more work for data preparation. You must use Keras package for creating your neural network model. Using neuralnet or other packages will NOT be acceptable.**
4. If your model has hyper-parameters, you must tune its hyper-parameters using cross-validation. Either auto-tune with caret or use a manual tune grid for each parameter (similar to what we did with tfruns for neural network). **You will lose some points if your model is not tuned.**
5. **You must compare the performance of your best model against a simple benchmark.** Model Benchmarking will be covered in Module 15 (lecture on responsible AI)
6. Use accuracy and AUC (Area under ROC curve) [for classification] and RMSE [for regression] on the validation data to compare multiple models, then report the performance of the best model on the test data
7. For Random Forest model, make sure that you report the variable importance and their ranking in predicting the outcome.
8. If you have a categorical variable with too many levels, consider using categorical embedding as will be discussed in week 13. If you have a textual feature which you think might be useful in predicting your outcome variable, then consider using word embedding as will be discussed in week 13 or tf_idf vectors as described in assignment 4.
9. With the exception of (video, voice, or images) Do NOT remove variable from your model unless you have a justified reason to do so. Removing variables arbitrarily from the model without proper justification can affect its performance and lead to under-fitting.

10. If target variable in a classification problem is imbalanced, make sure that you address the class imbalance, using one of the techniques that is discussed in lecture 12 (will be posted later)

Good luck and please email me if you have any question.

-Ellie