Project Charter

Please Enter Your Group Number here

Please Enter Your Names here (alphabetically)

Thursday, September 7, 2023

# Overview

This is an introductory statement that you will need to modify for the project you are choosing. For example, I am gathering data for my Reddit project. In your case, the introduction should read from a business or research “interest” perspective. Derive one question that you can/want to answer (see red text below) from the dataset of your choosing from the [UCI Machine Learning Repository](https://archive-beta.ics.uci.edu/ml/datasets?f%5Binstances%5D=greater-than-thousand&f%5BnumAttributes%5D=ten-to-hundred&p%5Boffset%5D=0&p%5Blimit%5D=100&p%5BorderBy%5D=NumHits&p%5Border%5D=desc). The dataset should have more than 1000 feature vectors and should be a classification dataset (in case you like a regression data set, let me know). Please post questions on the Discussion board for Module 2 for this submission. Use appropriate citations/ references as and when needed

The introduction for that project is “

*Nearly 430 million people use Reddit around the world. The message board’s popularity has grown from 300 million, a whopping increase of 30% since 2018. Reddit is a forum-based social media that focuses on sub-communities, called subreddits, each with its independent following. The traditional question-answer-based posts are the key to communication. The subreddit in question, r/wallstreetbets (the board or subreddit, hereafter), boasts a following of 9.2 million users. The subreddit has always been touted as a black sheep of the Reddit universe as it uses foul language and crude mechanics for stock evaluations. However, the subreddit is famous for fostering (yet abusive) community and excellent Due Diligence discourse. Recent market manipulation efforts have put the spotlight on the subreddit and pushed them to the center of the congressional hearing. These recent events have created an open question of whether these subreddit posts and comments contain relevant financial information. In line with this, we plan to address the questions. As such, this paper addresses the following questions: Can Reddit posts and comments predict the stock and derivative pricing? Do the higher posting activity and higher sentiment predict higher trade volume and volatility?*

*A remarkable range of odd and profane messages are posted on the subreddit. In line with the Financial theory, we hypothesize that the posters are “noise traders.” We test this hypothesis with the financial theory and market manipulation theory. We achieve this by utilizing the disagreement perspective (Harris and Raviv 1993; Karpoff 1986) and realized volatility perspective (Andersen et al. 2003).*”

All four stages below need to be “Documented.” It is important to not only program but also to explain each step of the program and analysis so that the “Deployment” team will fully understand the ML program you wrote and the analysis/result you want to achieve.

1. In the first stage, you will submit this document.
2. In the second stage, you will perform the analysis by using traditional methods (pick one that best suits the analysis) with generative and nongenerative methods. And compare the results for different parameter sets.
3. In the third stage, you will analyze the same dataset using random forest and classification. And compare the results for different parameter sets.
4. In the last stage, we will compare 2 and 3 and tune the parameters of the ML algorithm so that we will achieve better results.

## Project Background and Description

Write relevant project details here. You can delete the blue boxes (example of how to write it and wording to use) below when submitting the document. This description should clearly explain the “Whys?”

1. What is so interesting about this problem?
2. What are the proposed benefits of the solution (why do we need to solve this)?
3. What problem type are we evaluating (classification? regression or something else?)

|  |  |
| --- | --- |
|  | The complexity of networks and customer interactions will continue to increase in the future with the expanded deployment of various devices and new services. As a result, it is essential to enhance intelligence in network operations areas, including network operations, planning, optimization, and customer behavior areas, focusing on enhancing the overall customer experience. The application of machine intelligence will be central to automating network operations and optimizing the experience for the users of present and future networks.  The objective of this Project Group is to define what customer and technical requirements, practices, and infrastructure will be required to enable Machine Learning and AI to improve network operations and the customer experience. |

## Project Scope

Explain what you will do in each step given below. The example list is given below. You can use It and modify it for your project. The method below should provide an answer to your question in section 1. This section should explain “Hows?” of the project

What methods are we evaluating? (generative, nongenerative, trees? hyperparameters?)

What is the deliverable?

What is out of scope?

The list below prompts what you will do in all stages (your to-do list). You can use the list below (keep the image as it is).

Figure Project Stage Map

1. Generative Methods Based Analysis of the dataset
   1. Traditional exploration and analysis to predict the “dependent” variable.
      1. Preliminary visual exploration and analysis of the data set (exploratory data analysis).
      2. Propose a possible outcome of the analysis. For example, For the breast cancer data set, you can write. “from the preliminary exploration of the dataset, we can see that variables X1 and X3 are correlated. ”
   2. Use gradient boost to evaluate the dataset. Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
      1. Compare the results of the n\_estimator and learning rate over your dataset and pick the best parameter set. Explain what this means for the dependent variable for your
   3. Use bagging to evaluate the dataset. Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
2. Repeat 1.1-1.3 for the Nongenerative methods.
3. Hyperparameter tuning and analysis will be performed based on the results of 1 and 2. (you can keep the line as it is but delete this sentence)

## Deliverables

1. General Deliverables
   1. Common approaches, tools, and sharable implementations in support of ML/AI-based solutions for your project dataset, including:
      1. Data definition and normalization practices (breast cancer data set in sklearn has something called DESCR. You can see it by printing bc.DESCR, your dataset from UCI will have something similar. Please explain the variables used in that data set and how you plan to use them in your analysis here).
      2. Feature definition and training models (breast cancer data set in sklearn has something called DESCR. You can see it by printing bc.DESCR, your dataset from UCI, will have something similar. Please explain the variables used in that data set and how you plan to use them in your analysis here).
      3. Processing pipelines and optimization methods (Describe the analysis method you plan to use for this problem. Describe how will the use of the ensemble methods benefit you in addition to that analysis).
2. Generative Methods Based Analysis of the dataset
   1. Traditional exploration and analysis to predict the “dependent” variable.
      1. Preliminary visual exploration and analysis of the data set (exploratory data analysis).
      2. Propose a possible outcome of the analysis. For example, For the breast cancer data set, you can write. “from the preliminary exploration of the dataset, we can see that variables X1 and X3 are correlated. ”
   2. Use gradient boost to evaluate the dataset. Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
      1. Compare the results of the n\_estimator and learning rate over your dataset and pick the best parameter set. Explain what this means for the dependent variable for your
   3. Use bagging to evaluate the dataset. Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
3. Repeat 1.1-1.3 for the Nongenerative methods.
4. Hyperparameter tuning and analysis will be performed based on the results of 1 and 2. (you can keep the line as it is but delete this sentence)

## Constraints and success criteria

1.4.1. The success of the ML model is evaluated based on two criteria. One is the accuracy of the model predictive scores. In classifications, the accuracy is measured with precision, recall and F1 score. In regression model, MSE, R-squared, p-value, AIC & BIC are essential accuracy metrics to evaluate the success of modeling. The second criteria are whether the model makes sense from the food science and wine industry’s perspective, i.e., the features selected as predictive variables should be meaningful and recognized by the industry. The deployed model should also have enough Generalizability, meaning that the model’s accuracy should also be high when applied to the population where the sample is drawn or other sampled wines. Above all, the model should be able to help wine businesses to success in terms of wine inventory selection from the whole-seller, shelf displays and promotions.

1.4.2. There are two main constraints of this project. Given that the target variable is numeric (3-8), the subjectivity of bin size for the classification would largely impact the accuracy score, i.e., dividing the quality score into 2 bins (good and bad) would be easier to achieve high accuracy than dividing into 5 bins. There are no standard criteria about quality scores categorization, adhering to most popular industry and business requirement is the rule. Second, the dataset includes data only up to 2009, which might be missing recent wine quality measures. Third, the end users of the deployment might not have access to all the wine features used in the model, then the accuracy of the prediction would be impacted as a result.