CS - GY 6923: MACHINE LEARNING - PROJECT PROPOSAL

SHARK TANK US DATA ANALYSIS

Problem Statement | Motivation behind the project:

We are trying to understand the factors that influence startup funding by going through the data of the most successful startup funding show, Shark Tank US. Moreover, we are trying to analyze factors that influence the decisions of sharks like Mark Cuban, Barbara, etc. to invest in companies. Also, we want to understand these sharks' biases and build a model that can predict whether a startup idea can get funding from any of the sharks.

We harbor entrepreneurial dreams and would love to start a company in the US. Knowing about these factors will greatly help us understand the funding scene in the US.

Dataset Selected:

Kaggle URL: https://www.kaggle.com/code/thirumani/shark-tank-us-data-analysis

We are planning to use the dataset that is available in the above mentioned Kaggle URL. The dataset is not preprocessed. We will include the stage of Data Preprocessing in our project.

To make it clear, there are a few notebooks that are available that have preprocessed the data for data analysis purposes. But, for our multi-class classification purpose, we have to use our very own preprocessing pipeline.

ML Models:

We are thinking of combining the below models in the early stage of the project:

1. Logistic regression
2. Support Vector Machine (SVM)
3. Random forest

using the ensemble methods to classify whether a startup will get a deal or not.

The reasons for choosing these three models are mentioned below:

- The Logistic Regression model is highly interpretable and will give us insight into the features, but it cannot draw complex decision boundaries.
- On the other hand, the Support Vector Machine (SVM) model can draw complex boundaries, but it takes longer to train and could overfit the noise.
- The Random Forests model is highly interpretable and does not overfit easily compared to decision trees. As a result, to achieve state-of-the-art accuracy, we will combine the predictive power of all 3 models using the ensemble model.

Accuracy/Error Measures:

We will be focusing on multi-class classification. Therefore, we will use the below measures:

- Precision
- Recall
- F1-score
- Accuracy
- Confusion Matrices

For error measures, we will use the individual error measures of each model.

Previous Results and Improvements:

The results are available, but they are based on a previous version of the dataset, which does not include the newly launched seasons. Still, we will use the 3 classification models along with the ensemble model to test the previous results.

Project Proposal Submitted By:

Subhrajit Dey (Net ID: sd5963) ***Raunak Choudhary (Net ID: rc5553)***