Week 7 Meeting Notes

1. Call to Order

A meeting was held at 33 Charles St East on 21 October 2022.

2. Attendees

1. Regan
2. Riley
3. Jared

3. Open Issues for Week 7

1. Review Regan's progress on machine learning models
2. Review Jared's automation of docker build with github actions
3. Review other planned features our project intends to implement (description, feasibility)
4. Discuss acceptance criteria for these features

1. Machine learning pipeline
    a. PROGRESS:
       In progress. Have experimented with several models which will be elaborated on below.
    b. Highlights:
       Machine Learning Pipeline
        1. Feature selection
        2. Input into regressor
        3. Grid search cross validation

       Model Discussions and Conclusion
        1. Why does Decision Tree perform poorly
           Decision Trees learn simple decision rules from the features and determine the best split by choosing the feature that allows for greater information gain or lower node impurity. However, if there is no stopping criteria, the tree will continue splitting until all children nodes are pure. This causes the tree to overfit the training set and it becomes more complex than necessary.
           Due to overfitting, the tree was not able to perform well for the first 3 variations (baseline, outlier removal, feature selection). Only when a limit was placed to the depth of the tree during hyperparameter tuning, the tree was forced to terminate early and its performance improved significantly.

        2. Why is Gradient Boosting the better tree-based model in theory
           The tree-based models used are Decision Tree, Random Forest, and Gradient Boosting.

Random Forest is an ensemble learning technique - it constructs a set of base models from the training set and makes predictions by aggregating the predictions made by each base model. For Random Forests, the base models are Decision Trees and each tree is generated using a random subset of the training set and feature set. The forest then averages the predictions from all trees.

Random Forest is more accurate than a single Decision Tree as it incorporates more diversity. It is also more efficient as searching among a subset of the feature set when determining the best split is much faster too. However, it does not perform as well as Gradient Boosting.

Gradient Boosting uses boosting on top of ensemble learning. It boosts a set of weak learners to a strong learner such that misclassifications made by weak learners are made important. In general, the distribution of the training set is changed adaptively so that weak learners will focus more on errors made by previous learners.

For Gradient Boosting, Decision Trees are used as the weak learners and are added one at a time, and fitted to correct the errors made by previous trees. The errors made can be captured by a cost function, which is to be minimized using the gradient descent algorithm (hence the name gradient boosting). Thus, Gradient Boosting is the better tree-based model.

3. Best model to predict gross revenue for Top IMDB Movies dataset

   From a theoretical standpoint, Gradient Boosting should perform better than Decision Tree and Random Forest due to reasons made in earlier discussions. It should also perform better than the basic linear regressor. Indeed, from our results, Gradient Boosting performed the best among the 4 models for all variations (baseline, outliers removal, feature selection, hyperparameter tuning).

   From our results, it is safe to say that Gradient Boosting is the best tree-based model to predict the gross revenue of top IMDB movies from this dataset. However, as there are other strong models that were not covered in the scope of this project (eg. Multi-Layer Perceptron, Support Vector Machine), we cannot be certain that Gradient Boosting is the best model in predicting gross revenue for top IMDB movies.

4. Feature importance.

   We will use a feature importance score to rank the importances of our features. The feature importance score tells us the Mean Decrease in Impurity (MDI).

   The MDI of a feature is the sum of the number of splits across all trees where the feature is used, weighted by the proportion of samples at the split, and averaged by the number of trees. In other words, the higher the MDI, the better the feature is at reducing the tree's impurity.

We will use Gradient Boosting Regressor to determine feature importances as it performs the best among the baseline models. Only the top 30 features will be used in the models.

5. Top features to predict gross revenue for Top IMDB Movies Dataset
   Based on the feature importance ranking, the top 5 features are:
   1. Number of users who voted
   2. Budget
   3. Number of critic reviews
   4. Whether the movie is of the genre 'Family'
   5. IMDB score

   However, it is not clear how these features actually affect the gross revenue. We only know that they are important in terms of reducing the impurity for our Gradient Boosting Regressor.
   Out of the top 3 features which had the highest correlation with the target variable gross from the correlation matrix, 2 of them appeared in the top 5 features (Number of votes by users and Number of critic reviews).

6. Conclusion
   Every step of the EDA is vital in ensuring that our model inputs are not nonsensical. This ensures that our model predictions are meaningful (Garbage In, Garbage Out principle).

2. Automation of docker build
   a. By creating github actions to automate the process of building the docker image and pushing the image to the docker hub repository.
   b. This would also set the foundation for us to set up automated tests before building and pushing the image to the docker hub repository.
3. Planned features
   a. Create a feature that allows the user to input data of key features that the machine learning model has determined to get prediction results of gross revenue