PHASE 4 ~ PROJECT

FOCUS:

Advanced Machine Learning….

PROBLEM STATEMENT

- All entertainment websites or online stores have a lot of items. It becomes challenging for the customer to select the right one. At this place, recommender systems comes into the picture and help the user to find the right item by minimizing the options.

OBJECTIVES

The business objectives:

- 1. To create a Collaborative Filtering based Movie Recommendation System. It provides top 5 recommendations to a user, based on their ratings of other movie.
- 2. Predict the rating that a user would give to a movie that he has not yet rated.
- 3. Minimize the difference between predicted and actual rating (RMSE and MAE).

DATA UNDERSTANDING

- The movies dataset has 9742 rows and 3 columns, datatypes are Int64, Object
- The ratings dataset has 100836 rows and 4 columns, datatypes are Float64, Int64
- The timestamp column was deleted, the focus of the project being the movie id, user id, genre, title and rating.
- A merge of the two dataset was performed resulting to: 100836 rows and 5 columns , datatypes include float64(1), int64(2), object(2).

- **DATA CLEANING:** The Data contained no missing values or duplicates.

EDA(EXPLORATORY DATA ANALYSIS)

- Univariate Analysis: Countplot of the ratings, Barplot of the top 5 genres,

Title, User id and movie id (The unique id’s as well as the top 5 ID’s and titles were checked concurrently)

- Bivariate Analysis: The relationship between two variables : bar plot to show the top 5 movies with the best combined score .

MODELING

- The surprise library was used, a python-based tool specifically designed for building and analyzing recommender systems.
- Regression Model.
- Some imports will be performed :

* Dataset, Reader(enables for the dataframe to be converted to a surprise dataframe.)

- **Prediction Algorithms:**

* KNN basic – Its used by the KNN Basic Model stating the matrix used which is the cosine similarity matrix, **the algorithm used is the KNNBasic algorithm.**

* SVD – Its used by the SVD Model, the algorithm that's applicable is **the SVD algorithm**, matrix factorization is used for this algorithm

* KNNBaseline - – Its used by the KNN Basic Model stating the matrix used which is the pearson_baseline similarity matrix, **the algorithm used is the KNNBaseline algorithm.**

- **Model selection:**

* train_test_split - used for splitting data into train_set and test_set

* Cross validate - computes some accuracy measures.

- **Other imports:** Accuracy(A metric that measures how often a machine learning model correctly predicts the outcome) and mean_squared_error(Measures the average squared difference between the predicted and the actual target values within a dataset)

MODEL EVALUATION

- **Evaluate models based on Accuracy using the measure RMSE and MAE . After, we will pick the best model to tune for better performance .**

MODEL	RMSE	MAE	INTERPRETATION
KNNBasic	1.0100	0.7773	Baseline model, higher errors.
KNNBaseline	0.9758	0.7397	Improved performance (bias correction).
SVD	0.6448	0.4995	Best accuracy (captures latent features).

CONCLUSIONS

- The focus is on building a movie recommendation system using user-user similarity and matrix factorization. These concepts can be applied to any user-item interaction system.
- Explored generating recommendations based on a similarity matrix and collaborative filtering techniques. Additionally, I attempted to predict movie ratings based on a user's past rating behavior and evaluated the accuracy using RMSE and MAE error metrics.
- There is significant scope for improvement, including experimenting with different techniques and exploring advanced ML/DL algorithms.