Steam Game Recommendation System

Team 5

1. Nipun Hedaoo
2. Riddhi Shah
3. Varshini Vankayalapati
4. Vishanth Surresh

Research Questions

1. Study and develop a game recommendation system that provides personalized recommendation to video gamers to purchase a new game.
2. Multiple recommendation model is developed using multiple algorithms and performance comparison is done between them
    a. Matrix Vectorization - ALS
    b. Item Item Recommendation - Cosine Similarity and Pearson Coefficient
3. Calculate the effectiveness of each model using Root Mean Squared Error [RMSE] and Time Taken to train the model.

Dataset Description

Dataset Size	9.0 MB
Total number of Attributes	5
Number of New Features Generated	5
Attributes	UserID, Steam_Game, Behaviour Name, Hours_played
Total number of Datapoints	200 k
Sparsity of Utility Matrix	99.69 %
Total number of Files	1
Type	Recommendation System
Kaggle Dataset Link	https://www.kaggle.com/datasets/tamber/steam-video-games

In general, only 15% of the Games are more famous and has more number of ratings when compared to all other Games.

Notebook & Github Repo Links

Google Colab Notebook Link	https://SteamGameRecommendation
Github Repo Link	https://github.com/VishanthSurresh/Big-Data-Project
Kaggle Dataset Link	https://www.kaggle.com/datasets/tamber/steam-video-games

Colab Link

Github Repo Link

Kaggle Dataset Link

Note: The Github repo is in private mode. Only people who have access can able to access

Technology Stack and Development

Language	Python
Big Data Framework	Pyspark
Development	Google Colab, Databricks
Integration	Github
Python Libraries	Pyspark, Pandas, Numpy, CSV
Algorithms	Cosine Similarity [KNN], Matrix Factorization [ALS]

Implementation

1. Handling Missing Values
2. Handling Duplicate Values
3. Feature Addition
4. Feature Selection
5. Exploratory Data Analysis
6. Hyperparameter Tuning
7. Training
8. Testing
9. Validation
10. Model Performance Comparison - ALS Vs KNN
11. Conclusion

Model Design

Alternating Least Squares [ALS]	**K - Nearest Neighbour [KNN]**
The model is implemented using **Spark.ml Library.**	There is no Inbuilt library for **K - Nearest Neighbour.**
Latent Factor based Collaborative filtering is used.	Implemented item - **Item collaborative** filtering.
Hyperparameter tuning and Cross Validation is done, to determine best model parameters.	The model is implemented by finding **Cosine similarity and Pearson Coefficient distance.**

Model Implementation

ALS

- The model is developed using pyspark.ml and hyperparameter tuning is done.
- In ALS, data is directly splitted into train and test using random.split.
- Predicted Ratings for the test data.
- Found top 5 games for each user.
- In addition, displayed top 5 games for a specific user as well.

KNN

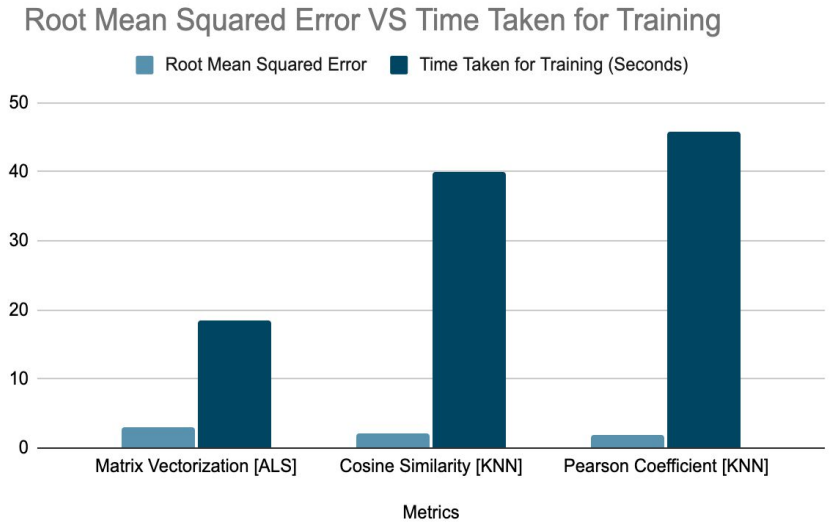
- Used Cosine similarity and Pearson Coefficient for each item-item combination.
- Model recommends both top 5 nearest games and games for a user based on similarity distance.
- Obtained the best model with low RMSE value in KNN using Pearson Coefficient Distance.

CPU Configuration

CPU Model	79
CPU Family	6
Model Name	Intel ® Xenon ® CPU @2.20 GHz
CPU Cores	1
Cache Alignment	64

Matrix Vectorization Vs Cosine Similarity Vs Pearson Coefficient

Metrics	Matrix Vectorization [ALS]	Cosine Similarity [KNN]	Pearson Coefficient [KNN]
Root Mean Squared Error	3.0166390	2.0756390	1.9758699
Time Taken for Training (Seconds)	18.57223	40.00675	45.87998



Model Comparison

Performance:	**ALS** pyspark model is way **faster** than **KNN** [Cosine Similarity and Pearson Coefficient]
Scalability:	While developing **KNN** model we faced lot of **scalability issues,** while ALS doesn't
RMSE:	The results of **KNN** is better than ALS. In KNN **Pearson Coefficient** model gave **better** results compared to Cosine Similarity

Model Evaluation

After considering multiple parameters, we found that the Pearson Coefficient model outperformed the other two models. However, the training time required for the Pearson Coefficient model was longer compared to the Cosine Similarity model and ALS. Despite this, the evaluation results showed that the Pearson Coefficient model was better.

Cosine Similarity and Pearson Coefficient results are almost same, only the key difference between the two models is that the Pearson Coefficient model subtracts the mean user game ratings from the rating data before calculating the similarity score, whereas the Cosine Similarity model does not. That's why Pearson Coefficient results are slightly better than Cosine Similarity.

Model Results

```
# User's ALS Recommendations:
# Joining recommended movies for the given user_id with anime dataframe to display recommendations in more readable format
int_df = nrecommendations.join(Games, on='USER_ID').filter('user_id = 151603712')
int_df = int_df.distinct()
int_df.count()
int_df.show()
int_df = int_df.toPandas()
d[3881]
int_df = int_df.sort_values(['rating'], ascending=[False])
int_df['Game_Name'] = int_df['GAME_ID'].map(d)
int_df
```

```
+---------+-------+--------+
|  USER_ID|GAME_ID|  rating|
+---------+-------+--------+
|151603712|   1170| 5.77904|
|151603712|   1033|5.362438|
|151603712|    518|5.444579|
|151603712|   2850|8.265933|
|151603712|   1905|5.479837|
+---------+-------+--------+
```

	USER_ID	GAME_ID	rating	Game_Name
3	151603712	2850	8.265933	NBA 2K14
0	151603712	1170	5.779040	Death Rally
4	151603712	1905	5.479837	Gems of War
2	151603712	518	5.444579	Binary Domain
1	151603712	1033	5.362438	Crypt of the NecroDancer

ALS Results

```
Steam Games Details shown below

Overall Avg Rating by user for games 151603712 is 0.5625

Top N Recommended games similar to user - 151603712 is shown below
```

```
[31] nw = newfeature2.filter(newfeature2['USER_ID']== 151603712)
     nw.show(5)
```

```
+--------------------+---------+--------------+------------+----------+-----------+------------------+------+
|          Steam_Game|  USER_ID|Behaviour_Name|Hours_played|prev_value|new_feature|    mean_Hourplayed|rating|
+--------------------+---------+--------------+------------+----------+-----------+------------------+------+
|The Elder Scrolls...|151603712|          play|       273.0|  purchase|          2| 86.10581818596883|     5|
|           Fallout 4|151603712|          play|        87.0|  purchase|          2|  61.2034090954641|     4|
|              Spore|151603712|          play|        14.9|  purchase|          2|25.605970128036258|     2|
|   Fallout New Vegas|151603712|          play|        12.1|  purchase|          2| 42.09011299998662|     2|
|       Left 4 Dead 2|151603712|          play|         8.9|  purchase|          2| 34.91489141701019|     2|
+--------------------+---------+--------------+------------+----------+-----------+------------------+------+
only showing top 5 rows
```

```
[32] topNrecommender.show(5)
```

```
+---------+--------------------+
|  USER_ID|          Steam_Game|
+---------+--------------------+
|151603712|The Elder Scrolls...|
|151603712|           Fallout 4|
|151603712|              Spore|
|151603712|   Fallout New Vegas|
|151603712|       Left 4 Dead 2|
+---------+--------------------+
only showing top 5 rows
```

KNN Results

Q & A

Thank you!