Movie Recommendation System

Matthew Gulbin

Business Proposal

Our company wants to provide the best movie recommendation system for content streaming services.

Our company will be training a model based on data sourced from MovieLens

Model 1: based on a smaller dataset with 100k reviews

Model 2: based on the full dataset with 33M reviews

Model 1: 100k Reviews

Data Summary

100000 reviews

- 600 users
- 9000 movies
- between 1995 - 2023

Each review has:

- A user ID, anonymized
- A movie ID
- A rating on a 5 star scale, with half star increments

Baseline Model

Created an SVD model with default parameters using surprise

Results:

RMSE: 0.88

FCP: 0.65

Hyperparameter Tuning

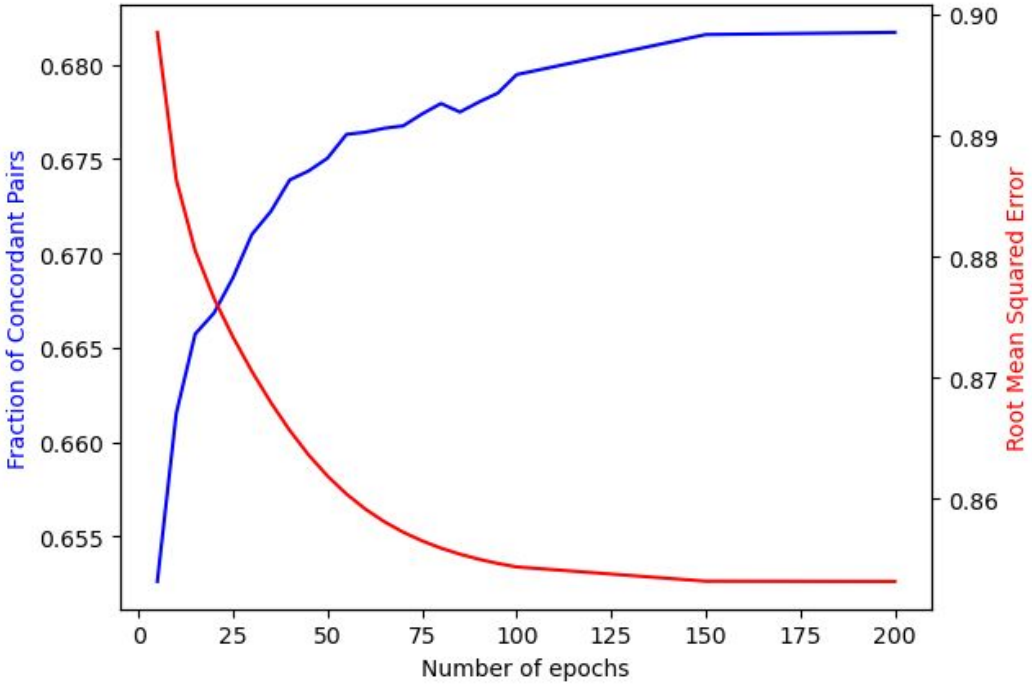
Hyperparameter tuning was focused on maximizing FCP

Found that the optimal parameters are as follows:

- *n_factors: 200*

- *n_epochs: 150*

- *regParam: .1*

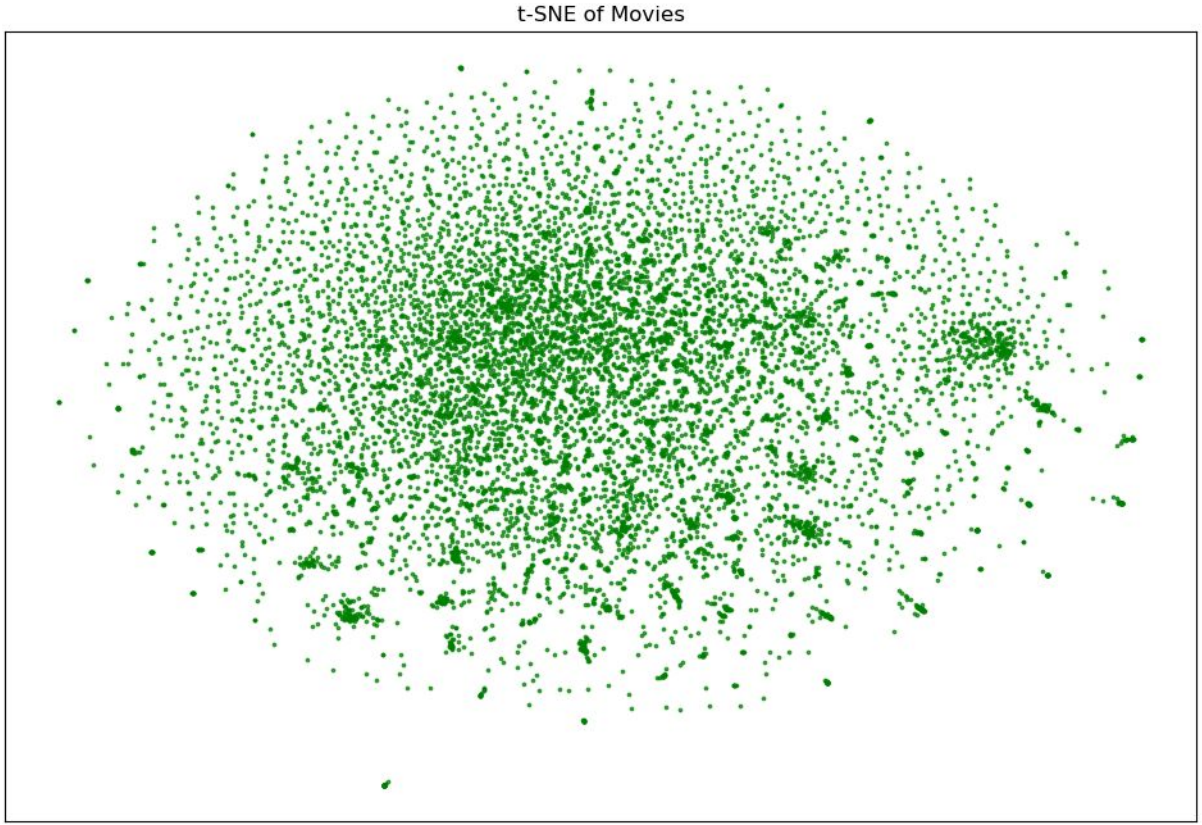


Final Model

- RMSE: 0.85
- FCP: 0.68

Created a t-SNE for each movie in the dataset

Closely clustered points represent movies with similar rating patterns



Model 2: Big Data

Data Summary

33832162 reviews

- 330975 users
- 86537 movies

Used Databricks and Pyspark to process and train models based on this large dataset.

Modeling and Hyperparameter Tuning

Baseline: Created an ALS model with default parameters

- ***Results: RMSE = 3.5***

Hyperparameter Tuning: Ran a cross validation on a small subset of the data, found that the optimal parameters are as follows: ***alpha = 0.5, maxIter = 10, rank = 15, regParam = 0.3***

Final Model: Trained on the full dataset

- **Results: RMSE = 0.91, FCP = 0.55**

Details

Created at	Mar 24, 2025, 04:13 PM
Created by	mjgulbin13@gmail.com
Experiment ID	2073136731611753
Status	Finished
Run ID	47563f973e1d477e927ac1410ee46ce5
Duration	20.9min
Datasets used	dataset (a81d1b76) Train
Tags	estimator_class: pyspark.ml.tuning.CrossValida... estimator_name: CrossValidator
Source	project
Logged models	—
Registered models	—

Parameters (15)

Search parameters

Parameter	Value
RegressionEvaluator.labelCol	rating
RegressionEvaluator.metricName	rmse
RegressionEvaluator.predictionCol	prediction
RegressionEvaluator.throughOrigin	False
best_ALS.alpha	0.5
best_ALS.maxIter	10
best_ALS.rank	15
best_ALS.regParam	0.3
collectSubModels	False
estimator	ALS
evaluator	RegressionEvaluator
foldCol	
numFolds	3
parallelism	1
seed	-3087263388934076904

Next Steps

- Continue hyperparameter tuning on the big data model to achieve a higher FCP
- Investigate the clustering in the data

Thanks!