DSCI 553 – Competition Project

Model Description:

A hybrid recommendation system was built by following two approaches and then merging the predictions by taking a weighted average of the predicted ratings from the two approaches. Here are the details of the approaches:

Approach 1: Alternating Least Square (ALS) Matrix Factorization for Collaborative Filtering using Spark

Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices. One matrix can be seen as the user matrix where rows represent users and columns are latent factors. The other matrix is the item matrix where rows are latent factors and columns represent items.

Approach 2: Ensemble model using XGBoost

XGBoost (Extreme Gradient Boosting) is a boosting algorithm that is based on Gradient Descent. It uses boosted trees and parallel boosting which makes it fast and accurate. Features such as latitude and longitude of businesses and cool, number of fans, number of funny reviews and number of useful comments of users were selected for training and prediction. MinMaxScaler was used to scaled the numerical features.

Final Prediction: A weighted average of predicted ratings were taken with 90% weightage given to predictions from XGBoost Regression and 10% to predictions from ALS Matrix Factorization.

Output Files

- als_model: Fitted ALS model
- model: pickled file with fitted XGBoost model and other
- prediction: Final json file with predicted ratings for each user and business pair in test_review.json