MovieLens Capstone-JHW

John Wilson

6/17/2020

Executive Summary:

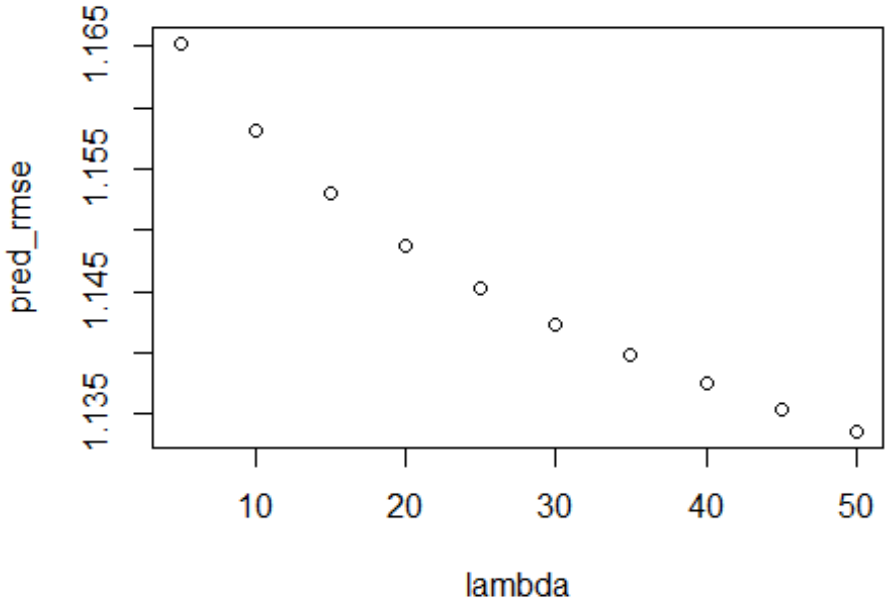
This project involves the MovieLens dataset that includes 10 million ratings for 10,000 movies utilizing 72,000 users. From its raw format, the data was cleaned up into tidy format over 6 columns (movieId, userId, rating, genres, timestamp, title) with the objective of being able to predict the rating. Furthermore, to confirm predictions, the data was divided into a working set and 10% going to a validation set to be utilized once a model was confirmed from the working set. After attempting several models to reduce the RMSE when compared to predictions, the GLM was selected which resulted in a RMSE of 1.03 on test data.The GLM model applied to the validation data resulted in a final RMSE of 1.05.

Technical Analysis:

The database included a multitude of genres and it was thought that by combining similar genres would later allow this to be used in a prediction model so the genres column was converted to factors and compressed to 50 categories via the fct_lump function. The working set (edx) was split into a test and training set to perform different variations of models, then validate each model. The first model attempted for predictions was a guess based on the average movie rating. Secondly movie and user effects were implemented to create a prediction. Timing effects were attempted but caused the system to freeze when trying to reduce timestamp to weeks. Hours would have been attempted also but would most definitely have frozen the program so the timestamp was negated in effect trials. From the effects modeling, the compressed genres did not improve prediction strength. Next model attempted was regularization with movie and user effects. Different methods of regression were tried on a 5000 lines sample of the training set as the full set was too large to do full regressions on. Also it was noted while running regressions that using the genres raw data caused errors but worked with the compressed factors for genres.

Results:

The results of the models were not as expected and further details will be provided. Increasing lambda on regularizing showed a reducing trend as seen in the plot below.



Further increases of lamda could have potentially reduced RMSE but at the current levels being so far above the average predictions it was not attempted for timesake. Surprisingly most of the models were at the average prediciton RMSE or higher which didn't make logical sense. Due to this, the second-best model was picked which was the GLM to apply to the validation.

Models	Mod_RMSE
Movie+User Eff	1.18
Movie+User+Genre Eff	1.18
Reg Movie+User Eff	1.13
Knn	1.10
Avg Pred	1.06
Loess	1.04
Ensemble	1.04
GLM	1.03

GLM applied to the validation set resulted in a RMSE of 1.05.

Model	Val_RMSE
Validation RMSE	1.05

Conclusion:

Predicting rating based upon the information presented in the dataset was a challenge but different methods were applied to get a sufficient model to predict accurately. RMSE was used as the grading of the 8 models that were produced. Due to the size of the data regressions had to be sampled in order to perform them which prevented the ensemble from having all the models included. Other limitations to the model were the vast spread of genres and no indications of movie length or awards won which could be helpful in prediction. Overall, it was a good exercise to implement skills and maneuvering of data to make models work. With more time, getting the qda/lda models to work and attempting more trials with lambda to see where the minimum of the model's limitations were could have been achieved. Nevertheless, With the current data and model setup the best achieved RMSE was 1.05 with GLM model.