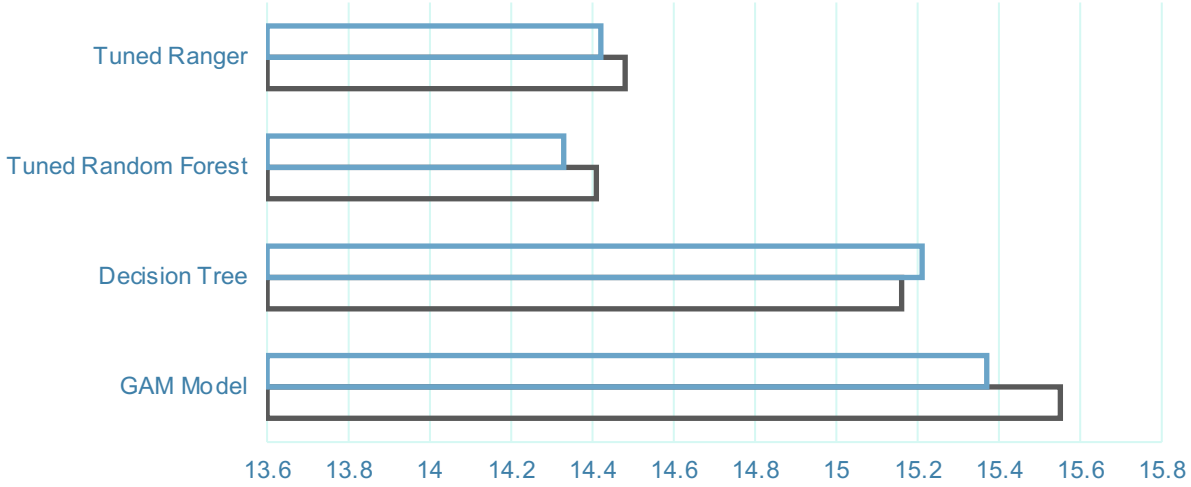
My thought is based on the Frequency of Genre

Zhenyu Wang (zw2847)



	GAM Model	Decision Tree	Tuned Random Forest	Tuned Ranger
RMSE in Kaggle	15.37	15.21	14.33	14.42
RMSE in test	15.55	15.16	14.41	14.48

My best model: **Random Forest Model** with tuned **mtry = 73**, **ntree = 1000**, and **count frequency** to create the top 700 genre types as dummy variables. The **Kaggle RMSE** is **14.3359**.

Methods

Feature selection:

1. Extracting unique genres as a list
2. Count the Frequency of the list
3. Based on the list to create dummies

Model:

1. GAM model by adding smoothing term
2. Decision Tree
3. Random Forest with tuned
4. Random Forest with Ranger tuned

Conclusion:

In this project, the key is dealing with genre variable, and the **random forest model** is a powerful method for handling such **high-dimensional data**.

Improvement:

Find a better way to select genre types, not all the genres contribute useful information to the model.