**Summary**

**Problem**: Build a machine learning prediction model for song popularity on Spotify using 10 song-based variables

**Performance measure:** RMSE will be used to evaluate the performance of the regression models

**Process**: Start small with simple linear models, before moving to more complex methods based on as much reasoning as possible. As each phase reveals general techniques such as scaling, they are re-implemented from the beginning.

**Results:** Stacking Ensemble using decision tree ensemble models and an SVR model with an rbf kernel produced the lowest RMSE score of approximately 10.3.

**Data Exploration and Processing**

* Dataset has 11 numerical variables, 10 independent and 1 dependent (song popularity)
* Dataset is small with only 453 observations suggesting more sophisticated splitting method might be useful for training models, especially for ensemble where multiple models are used
* A pair plot of the numeric variables reveals lack of linear relationships between the independent variables and the dependent variable
* This suggests linear models may have little predictive power.
* A correlation matrix confirms the lack of linearity.
* The ‘top genre’ feature was broken down using one-hot encoding as it marginally improved predictive power across most models.

**Modelling**

**Across Phases: Common Techniques**

* Cross validation was used for all models in this study due to the small size of the dataset and the unclear relationship between the independent and dependent variables
* Grid Search was used for all models in order to find to the optimal values for the hyperparameters. This was particularly useful for Support Vector Regression models where many hypermeters are available.
* The numeric variables were scaled for all models in this study. The decision was made during the phase 1 investigation which revealed a mean RMSE of 14 for unscaled data versus 11 for scaled data.

**Phase 1: Linear and Logistic Regression Models – Scaled or Unscaled, Feature Selection**

* Overall, linear outperformed logistic with minor differences between scores for linear, lasso and ridge (approx. 11 RMSE)
* To evaluate the linear relationship in the data, a variety of linear models were run on the complete training set with unscaled data. These include Linear, Lasso, ElasticNet and Ridge. The mean RMSE for the scaled data was 11 while the mean unscaled RMSE was 14, favouring scaling for this dataset.
* A smaller set of features with relatively higher correlation had a worse RMSE performance than the complete set of features, favouring using the dataset in full.
* Similar results were observed for the Logistic regression regarding scaled vs. unscaled, and complete vs. selective feature selection.

**Phase 2: Decision Trees**

* The ensemble method was used for this phase, specifically the Random Forest Regressor and the Extra Trees Regressor as it produces better predictive than utilising only one decision tree, especially due to the complex nature of the dataset
* Both improved the RMSE by 1 unit (approx. 10.3 RMSE), despite different hypermeters from running Grid Search
* This might be due decision tree models being more suited to capturing non-linearity in the dataset

**Phase 3: Support Vector Machines**

* We implemented a variety of Support Vector Regressions due to their ability handle non-linearity better than linear regressions by using a certain threshold for tolerating smaller errors
* Grid Search correctly found the non-linear Rbf kernel to have the lowest RMSE out of all SVR models (approx. 10.6 RMSE) but no apparent improvement on the decision tree models

**Phase 4: Ensemble**

* Ensemble were investigated in order to fit the data using multiple regressors as individual models were not capturing the data on their sufficiently
* A voting regressor and a bagging regressor were used.
* Voting significantly outperformed bagging (~10.4 versus 11.4 RMSE)
* **NEED TO RE-RUN BAGGING (it has not been fitted with a selection of models for parameter selection)**
* The stacking regressor delivered the tied-lowest score of the study (~10.3), matched by the ExtraTreesRegressor on its own.
* Three base models were used, namely, the ExtraTreesRegressor, the RandomForestRegressor and the SVR using the rbf kernel. The meta learner was a basic linear regression.

**Conclusion**

* As illustrated by the diagram below, the stacking regressors is the best performing model across all the training samples. We therefore chose to create our predictions using this model over the ExtraTreesRegressor.

**Other models**

* K-nearest neighbours (approx. 10.5 RMSE)
  + Good on its own, but not in an ensemble
* Gradient Boosting model (approx. 11.7 RMSE)
* AdaBoost Regression model (approx. 10.9 RMSE)