Predicting Movie Ratings using Various ML Models

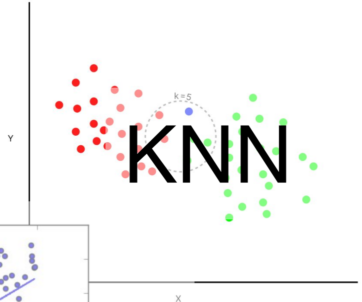
Team ROB: Ria, Omar, Bryce

Rating:
75/100

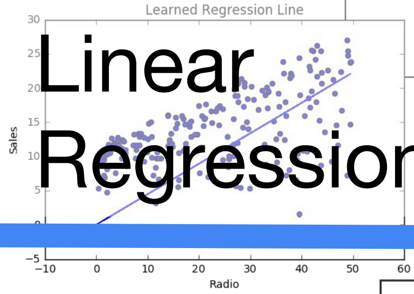

IMDB Database
10179 Movies
- Title
- Release Date
- Genre
- Description
- Status
- Language
- Budget
- Revenue
- Country
X-MEN
DAYS OF FUTURE PAST
Rating:
75/100

IMDB Database
10179 Movies

- Title
- Release Date
- Genre
- Description
- Status
- Language
- Budget
- Revenue
- Country



KNN



Linear Regression

Random Forest

Rating: 75/100

Methods

K-Nearest Neighbors

- non-parametric method
- Parameters used; Release date, Genre, Original Language, Budget, Country
- Release date = year, month, summer, valentines, halloween, christmas, new year
- Algorithm: Remember all of the training data. For new points: find the k closest points in the training data. For regression: typically take the average of the labels.
- Note, ALL SAME WEIGHT

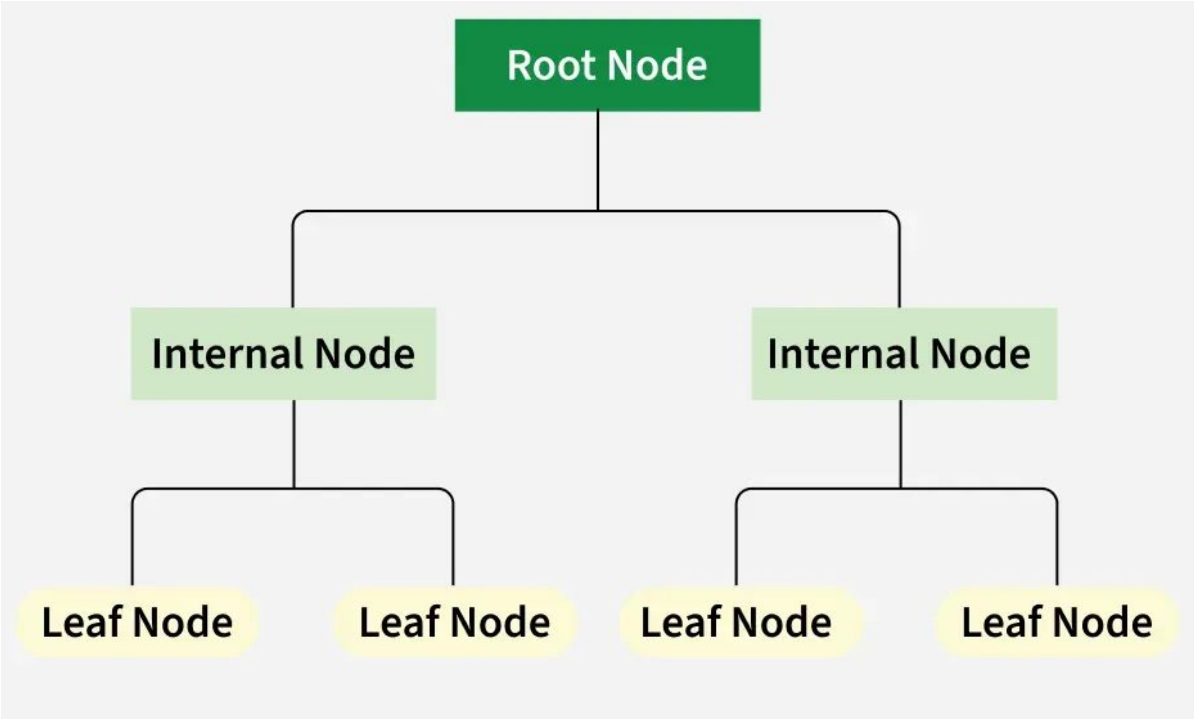
K-Nearest Neighbors

- Measuring accuracy: root-mean-square deviation
  - i = variable i
  - N = number of non-missing data points
  - xi = actual observations time series
  - \hat{xi} = estimated time series
- 80% training set 20% test set.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

Decision Tree

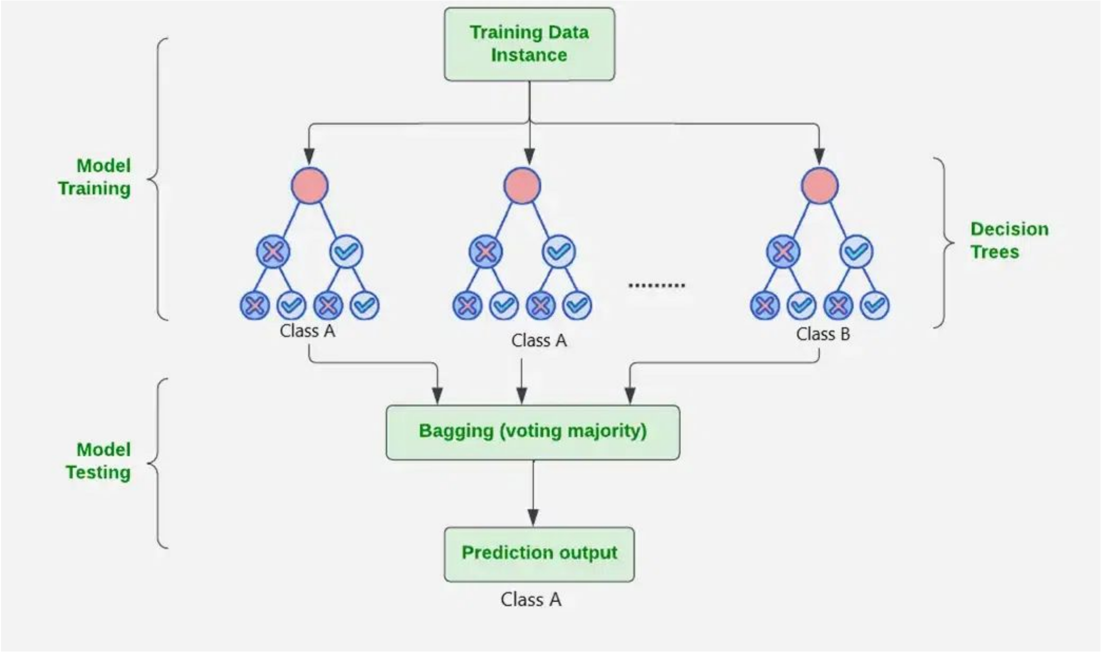
- Uses data to classify the outcome of a certain data point
- Splits into two options based on the parameters of previous node



GeeksForGeeks

Random Forest

- Use several different decision trees to
- Best to use when importance of datapoints
- Meant to avoid overfitting and reduce bias



GeeksForGeeks

Random Forest Parameters

- Profit (Revenue-Budget)
- Vectorize valuation of Overview
- Genre
- Country
- Language
- Classification vs Regression
  - Majority for classification
  - Average for regression

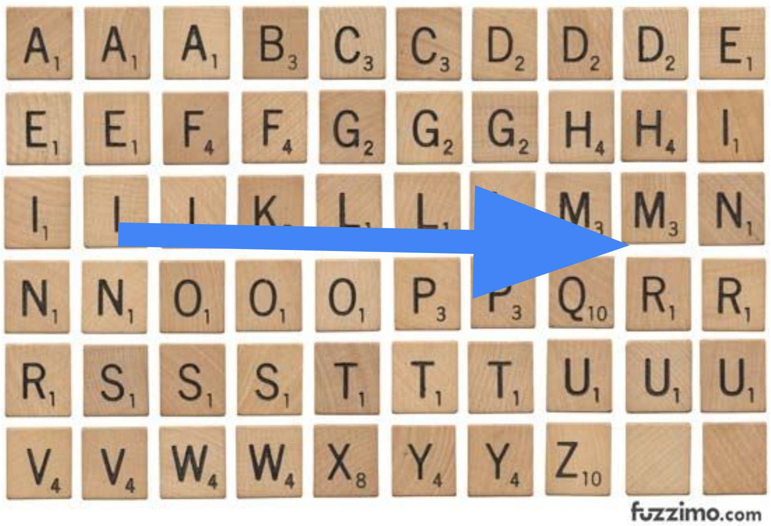
Linear Regression

Title, Release Date, Genre, Description, Status, Language, Budget, Revenue, Country

Linear Regression

Title, Release Date, Genre, Description, Status, Language, Budget, Revenue, Country

Text



Linguistic Aspects

Linear Regression

- Description vs Rating
- Revenue/Budget vs Rating
- All Features vs Rating*

Linear Regression

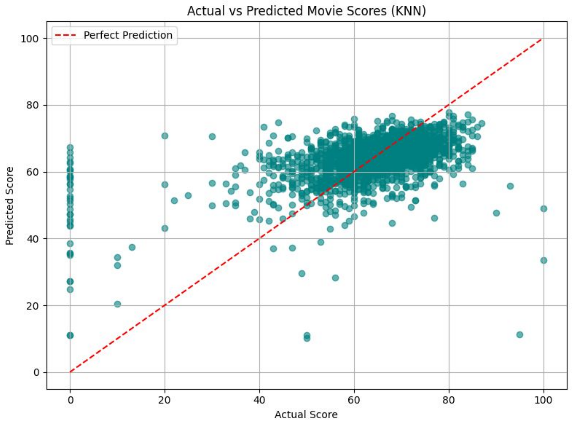
Linear Regression: MLE

- Let $\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$
- Model
  - $f_{\widehat{\boldsymbol{\theta}},\sigma}(y_i|\boldsymbol{x_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mu_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x_i})\right)^2}{2\sigma^2}\right)$
  - $\mu_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x_i}) = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x_i}$
- MLE
  - $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
  - $\sigma^2 = \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\right)$

Linear Regression

Results

K-Nearest Neighbors

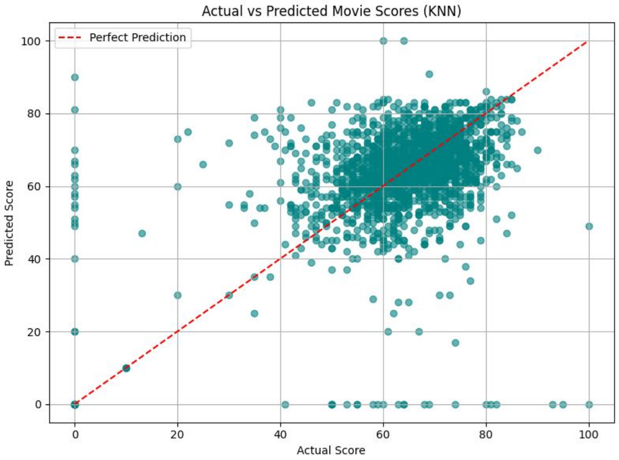


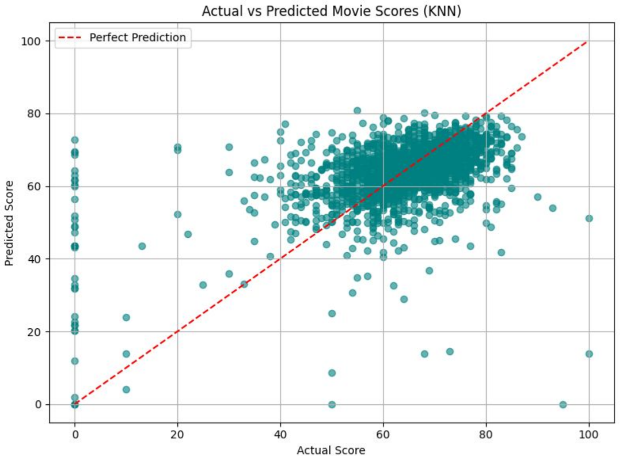
K=1,
RMSE: 13.99,

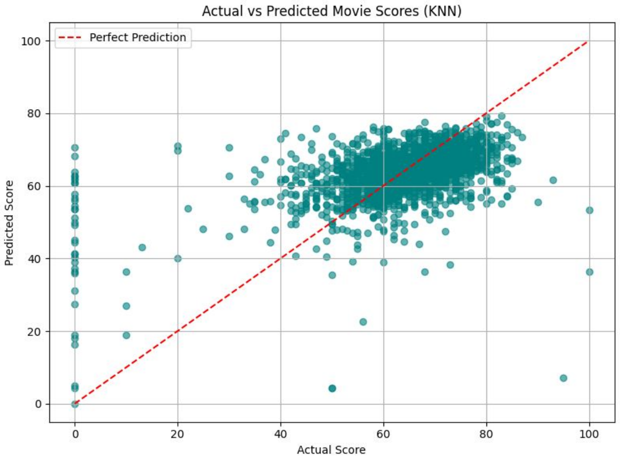
K=5,
11.43,

K=10,
11.05,

K=20
10.99

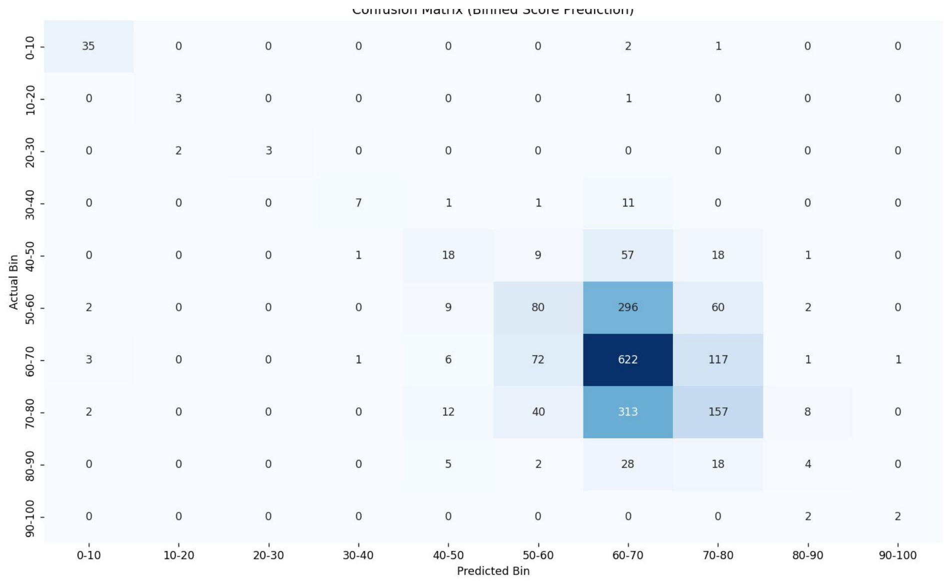


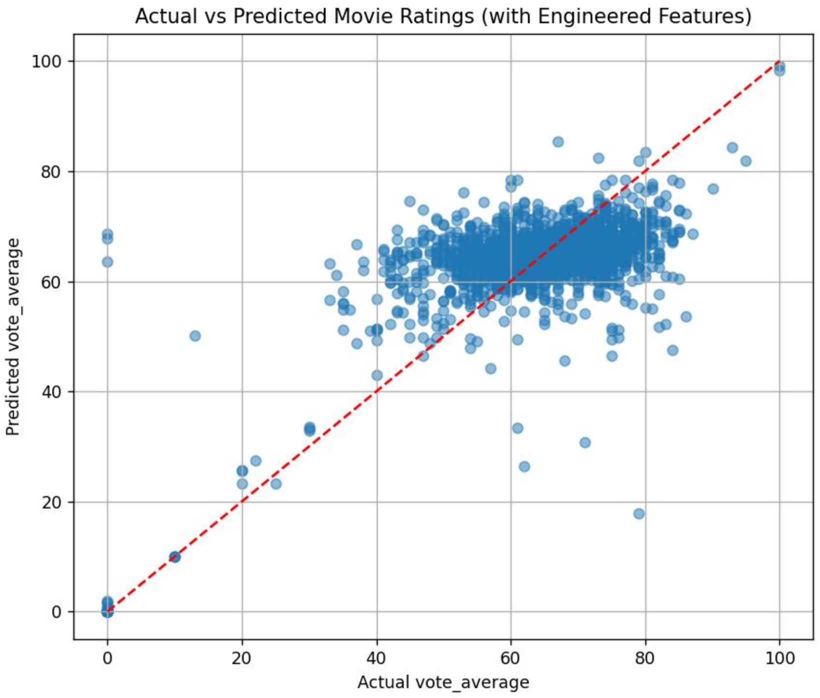




Random Forest

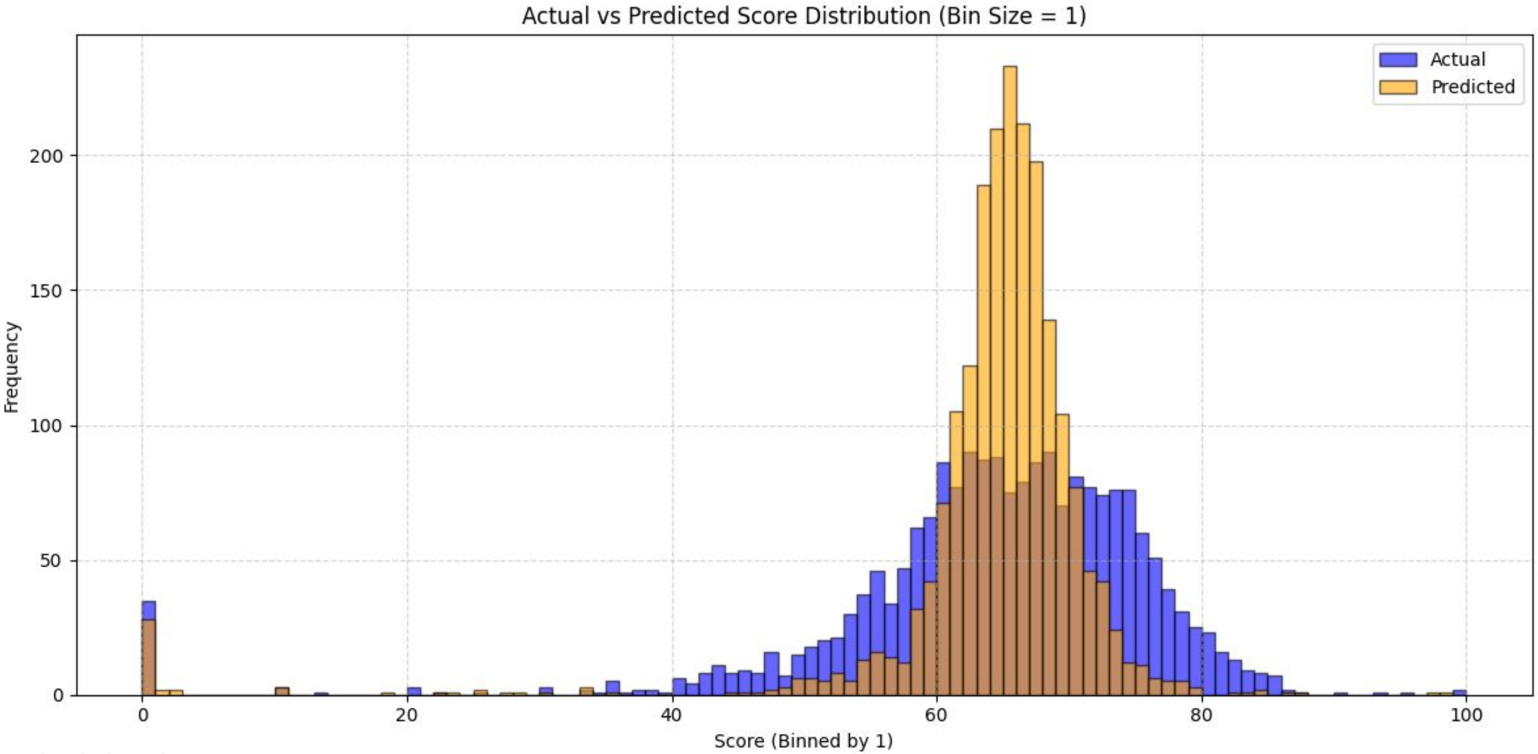
RMSE: 9.0952





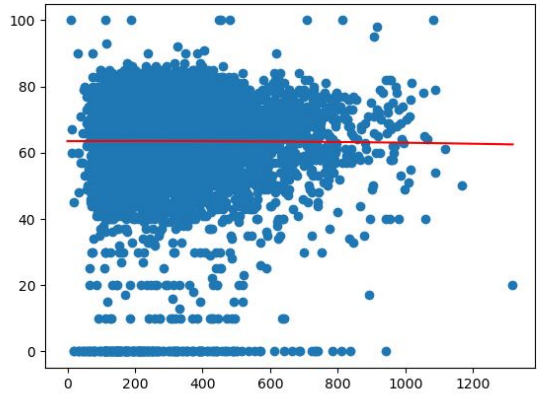
RMSE: 1.1171 bins

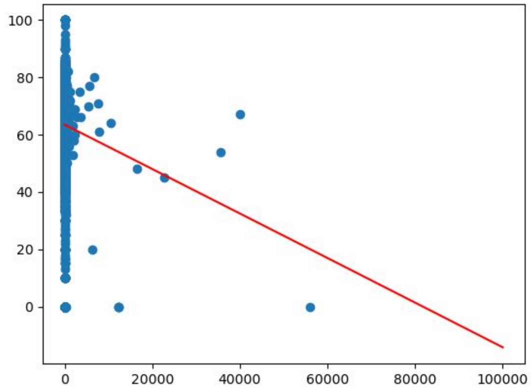
Random Forest (Overfitting)

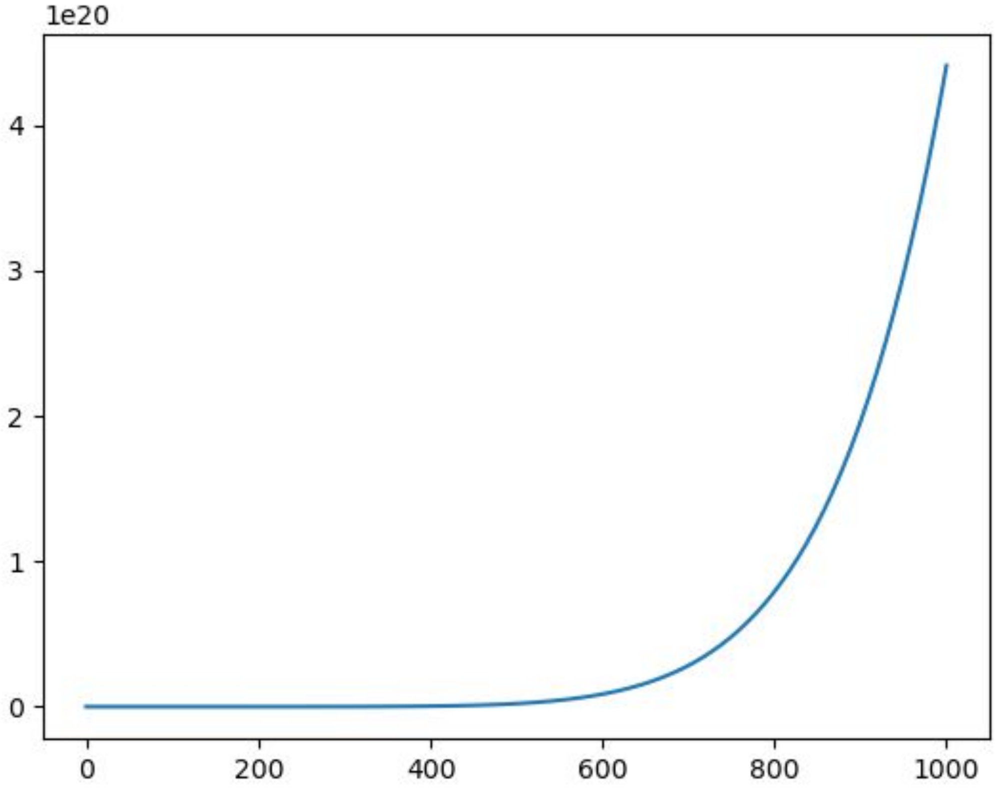


RMSE: 9.0952

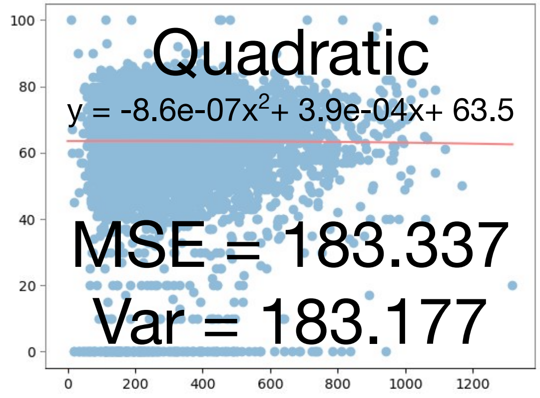
Linear Regression

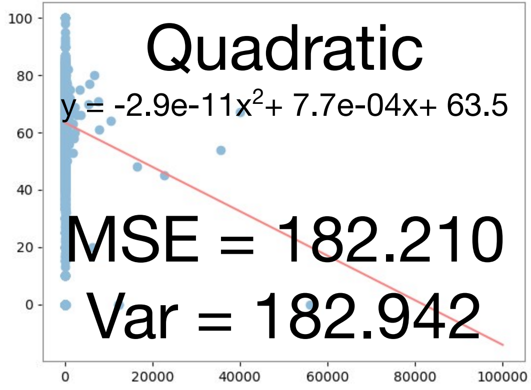


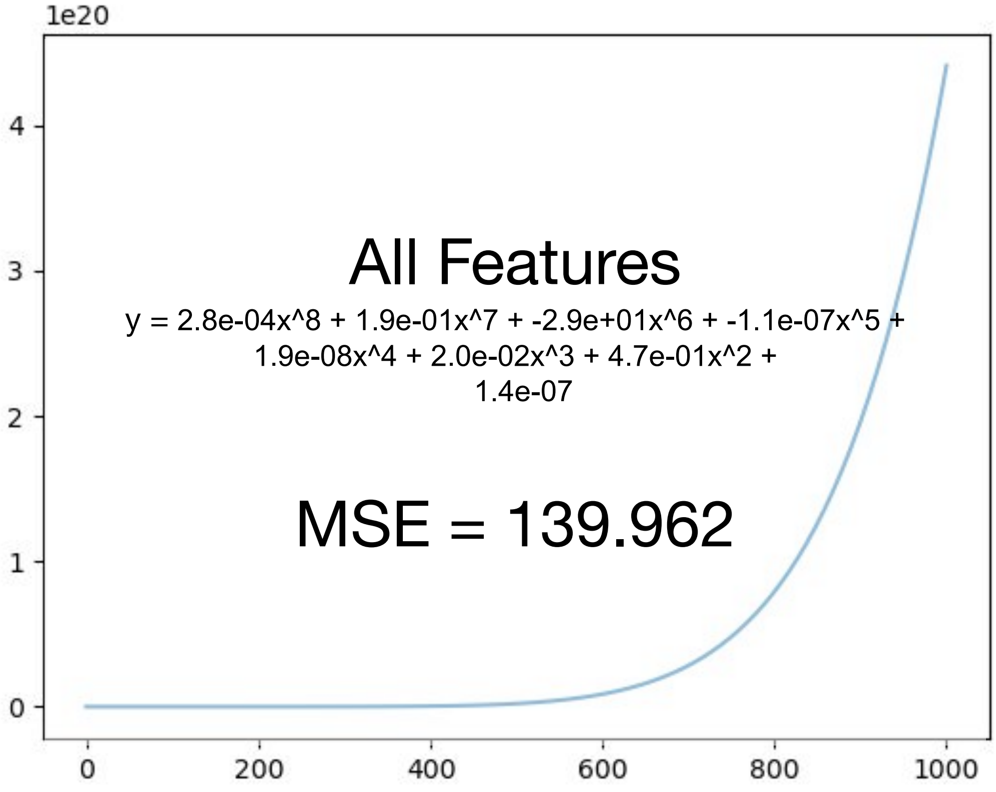




Linear Regression







Conclusions

Our approach was ambitious…

Random Forest is the GOAT out of each of our models

But generally, this dataset has a couple factors which makes it difficult to use for prediction!

→ missing cultural factors which contribute to a movie's rating: cast, director, marketing, etc

Thank You!