Video Game Sales Prediction

By:

Hari Krishna Rangineni

Likith Kumar Miryala

Sai Kiran Putta

Scope of the Project

- Models performance on the dataset

- Prove that feature generation from “Neural Networks” are superior!

Feature Set

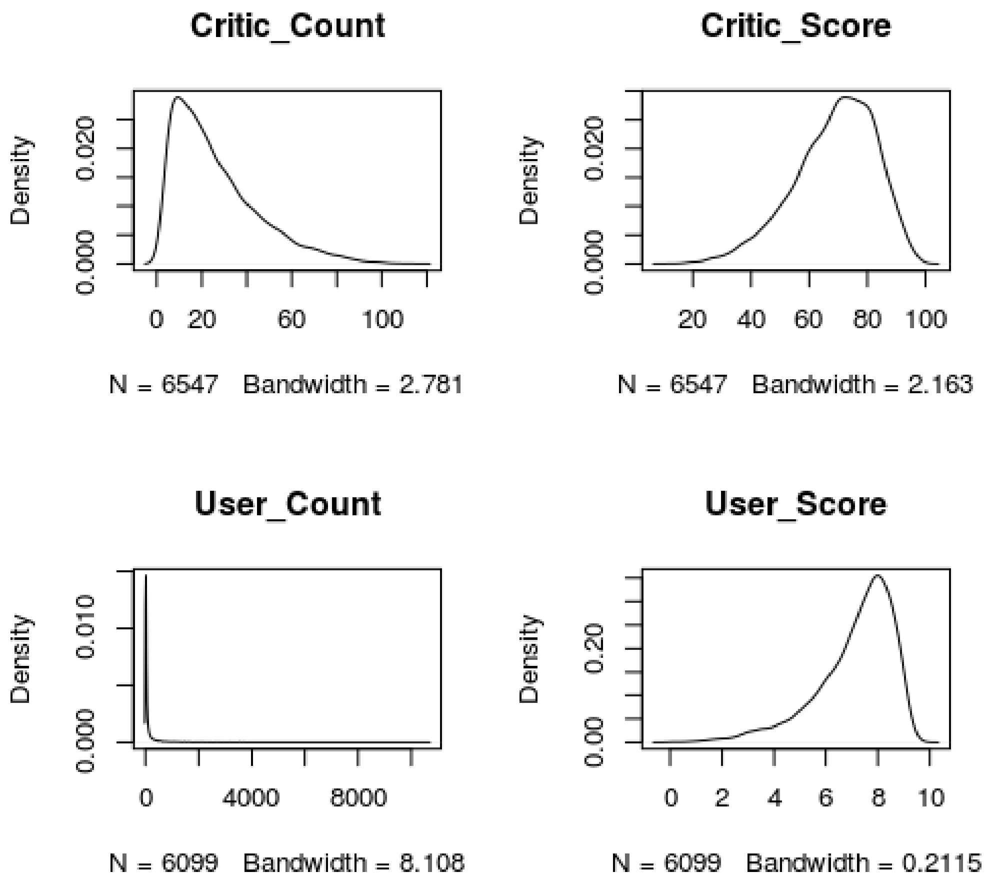
- Name
- Platform
- Year of Release
- Genre
- Critic Score
- User Score
- Critic Count
- User Count
- Global Sales
- Rating
- JP sales
- Other sales

Deleted Features:

- EU Sales
- Developer
- Publisher

Mean Vs Median for Imputation

- **Taking a Median is better than Mean since the distributions are skewed.**



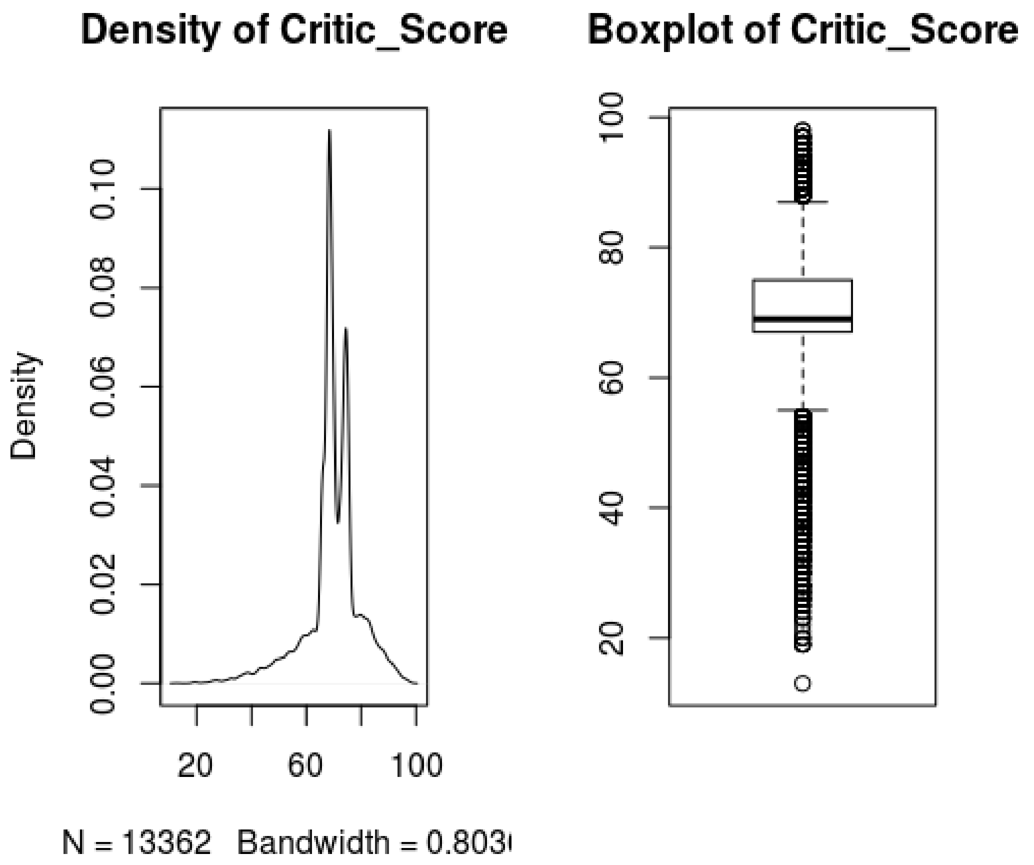
NA Imputation Strategy

- **Imputation by Grouping:**

Grouping by Genre and taking a median of feature.

Outliers In Data

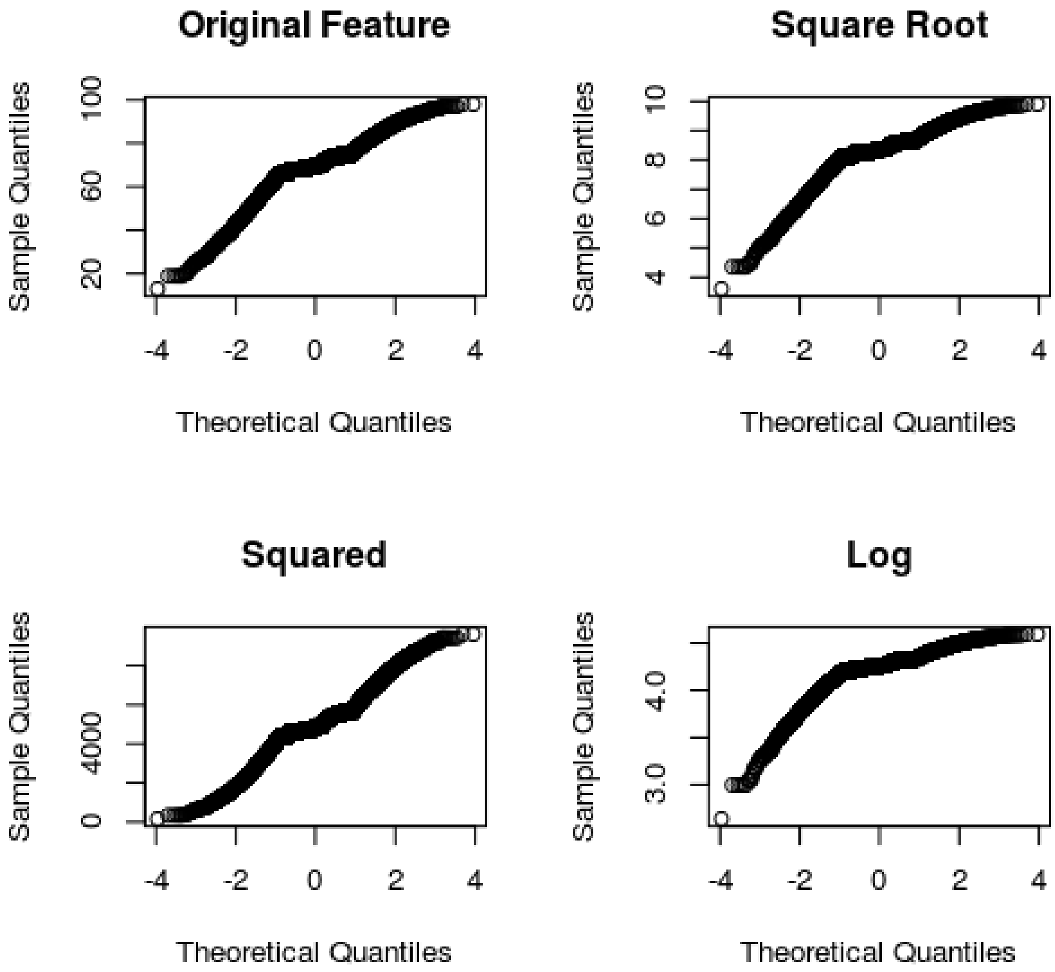
- ➢ If we go by boxplots, most of the data will get deleted.



Outliers In Data - 2

Our approach:

- **Do feature transformation first, then remove outliers.**



Feature Engineering.

- Length of Game Name
- Length of Publisher Name
- In Platform: X360 / XB / XOne -> Xbox

  PS / PS2 / PS3 / PS4 / PSP / PSV -> PS etc.
- Year by different periods

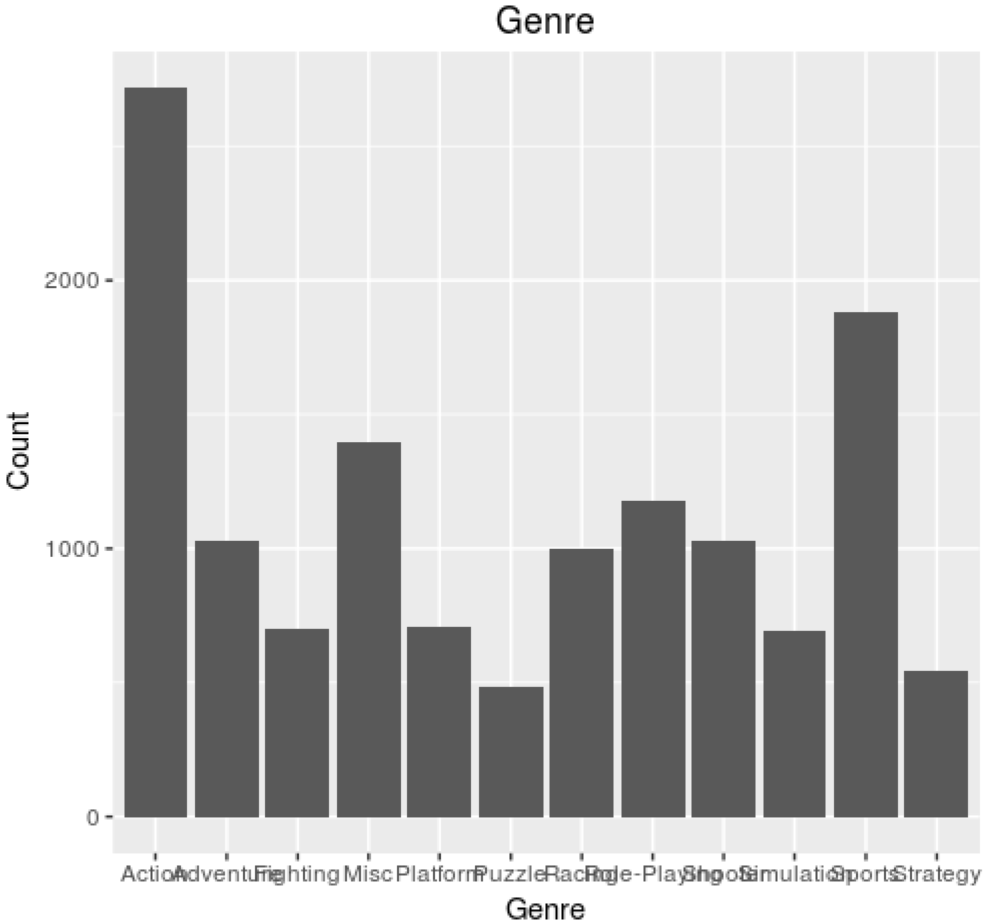
  YearI -> 1980 – 2015 by step of 5 years

  YearII -> 1980 – 2015 by step of 10 years

Exploratory Data Analysis.

Some Interesting Plots

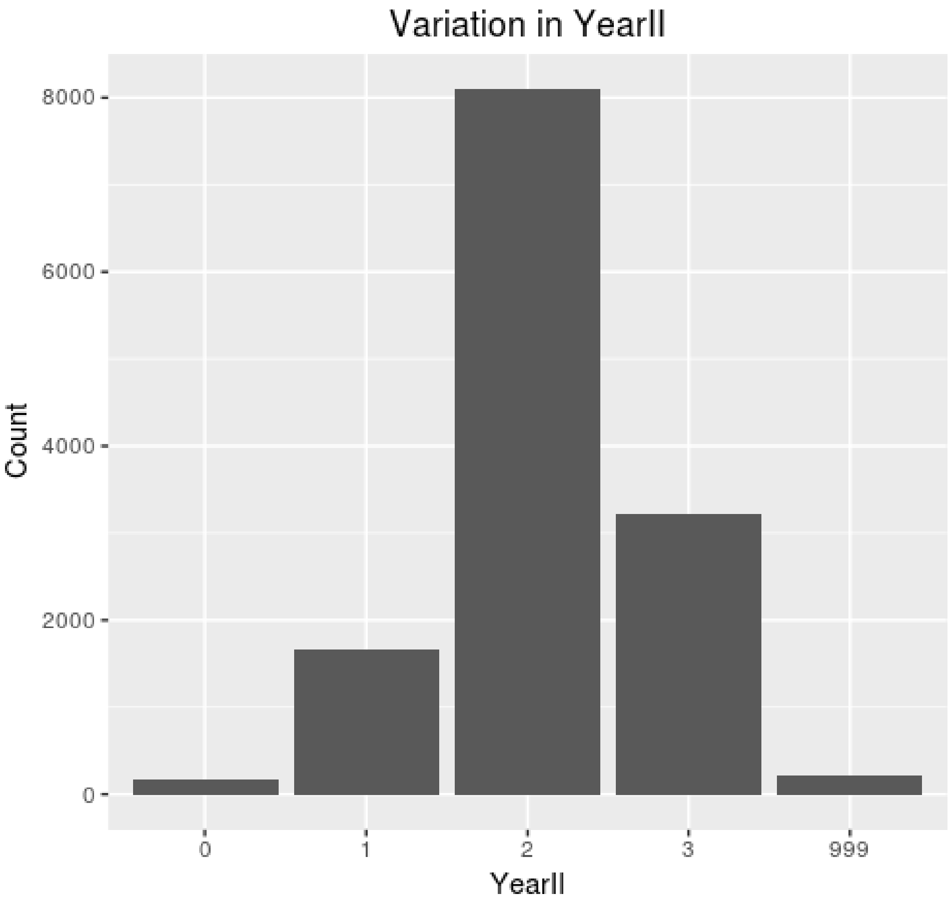
Inference: **More number of Action, Sports related games.**



Some Interesting Plots - 2

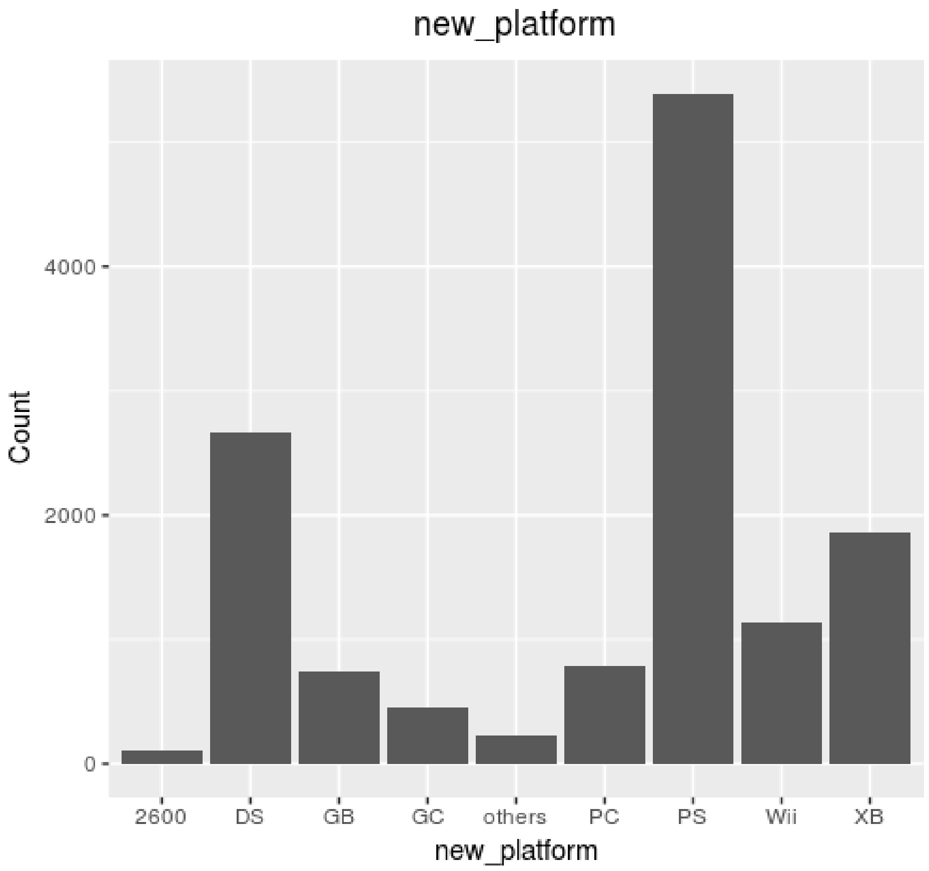
Inference: **More number of games were released in 2000 – 2010**

*** 999 is category where we don't know the year.**



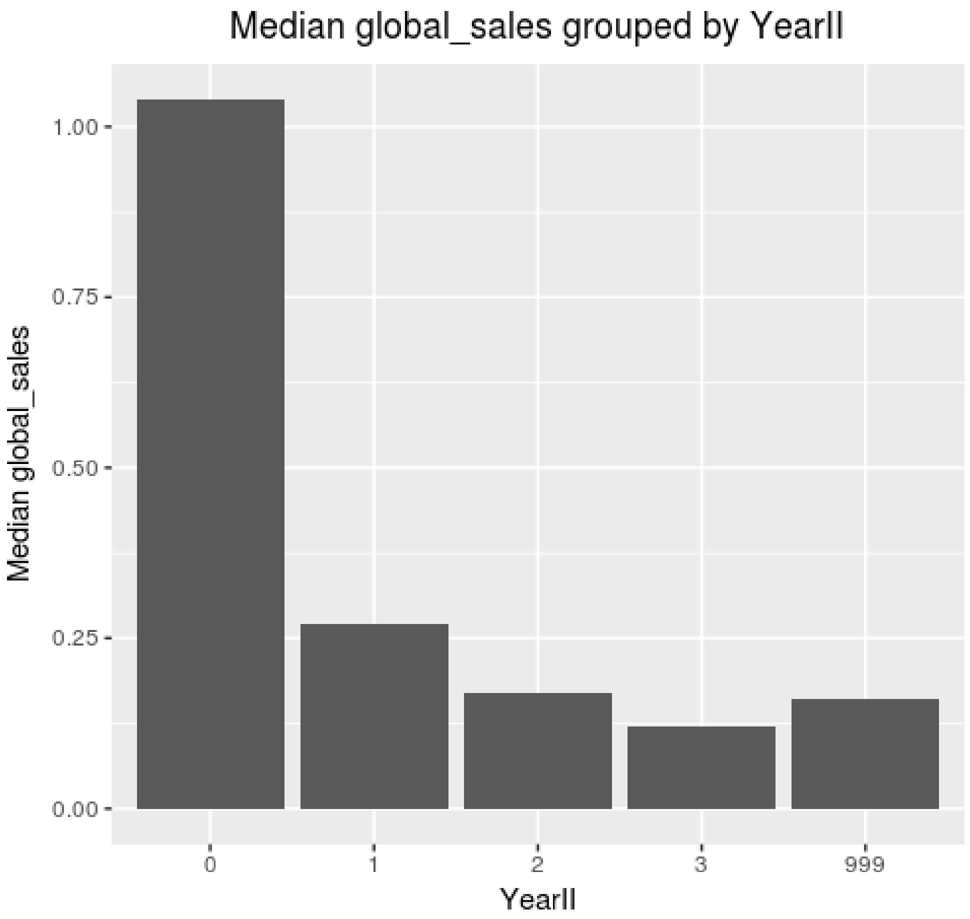
Some Interesting Plots - 3

Inference: **More number of games were released in PlayStation followed by Nintendo DS.**



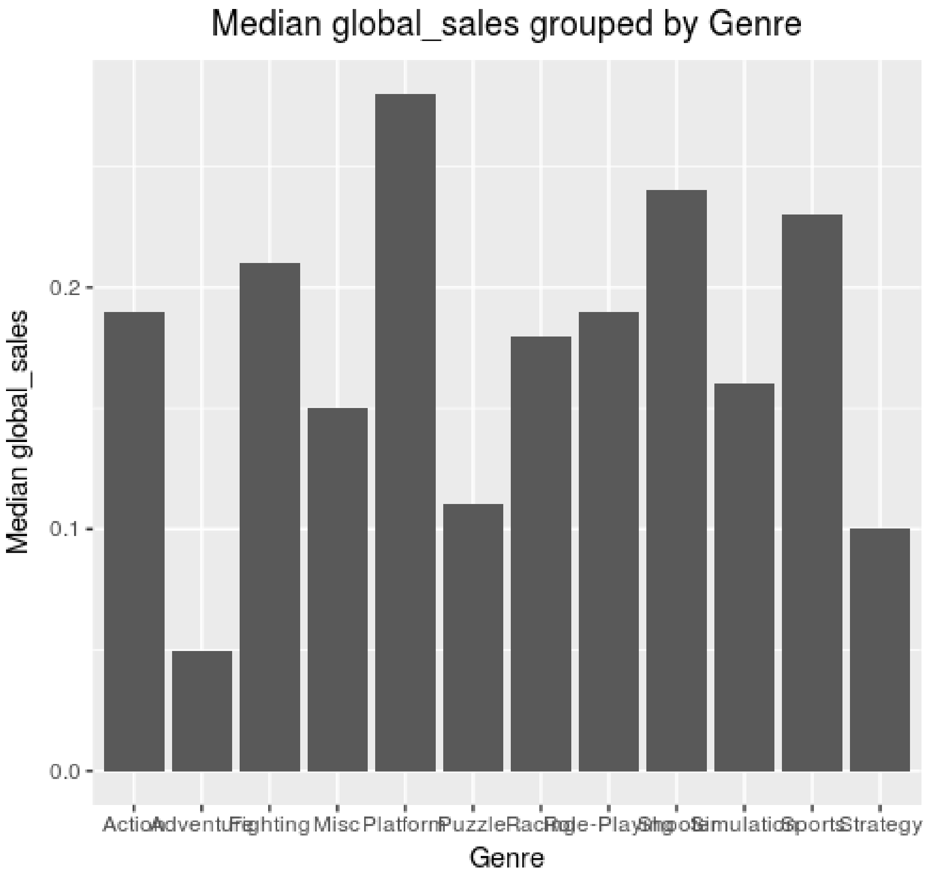
Some Interesting Plots - 4

Inference: Interestingly Global sales in 1980's is very high despite number of games in 2000 – 2010 are more.



Some Interesting Plots - 5

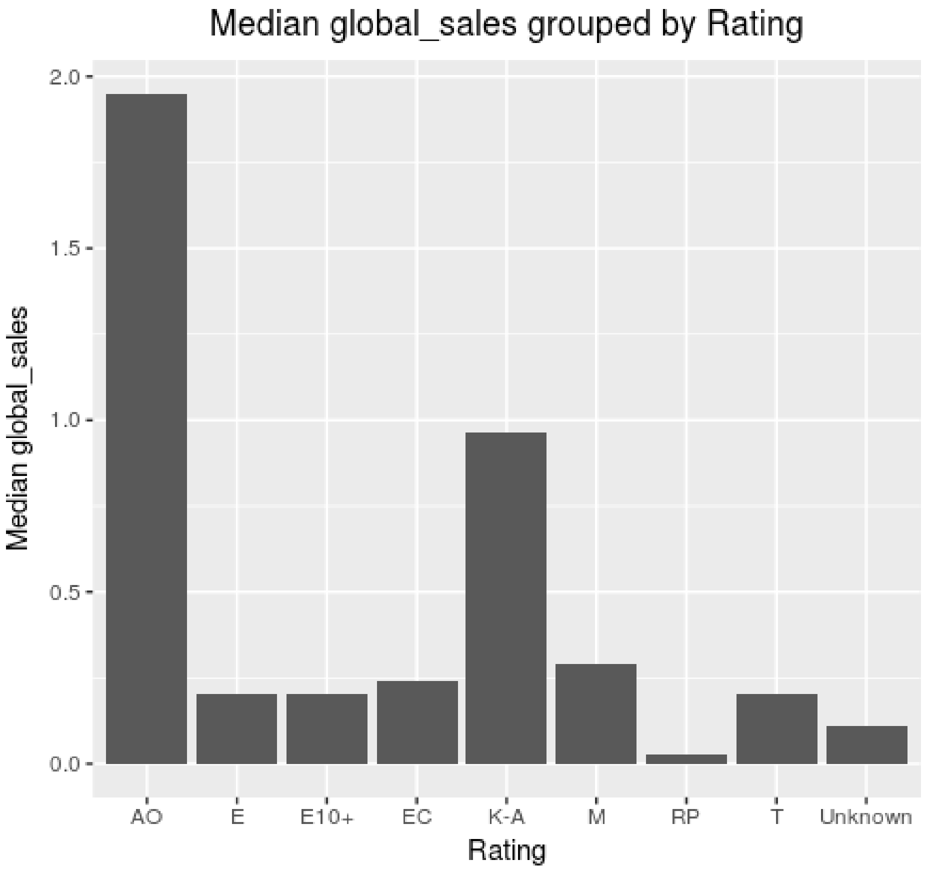
<u>Inference:</u> Global Sales are more in Platform, Shooter and Sports Genre in the increasing order.



Some Interesting Plots - 6

Inference: Global Sales of “Adult Only” rated games is significantly high.

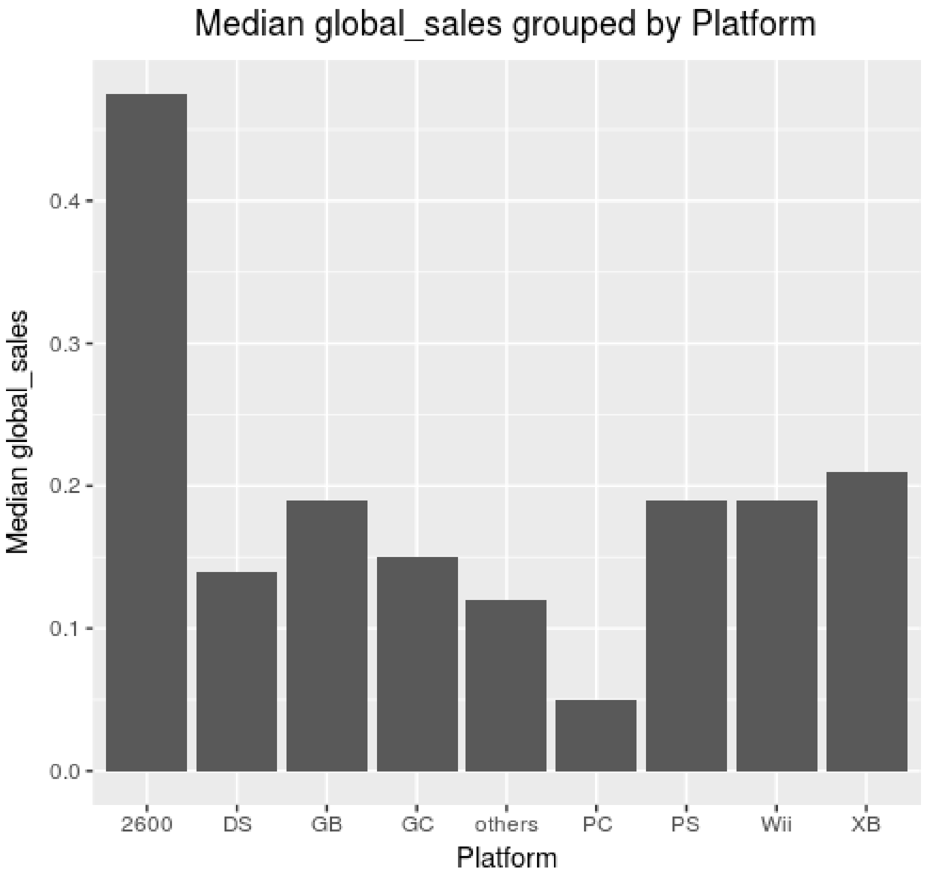
New Feature: If a game is AO or K-A, 1 else 0.



Some Interesting Plots - 7

Inference:

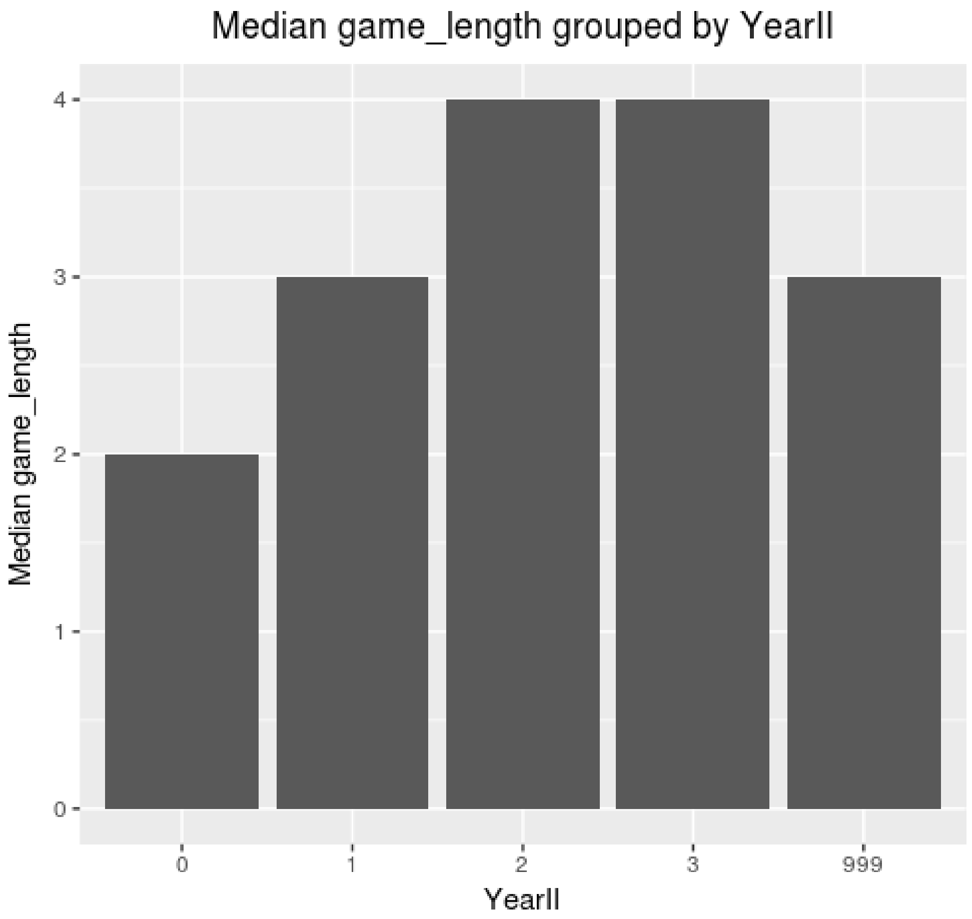
- ➢ **Global Sales of 2600 is high since the number of games were more during that era.**
- ➢ **In current competitors Xbox has high sales!**
- ➢**New Feature: If 2600/XB/PS 1 else 0.**
- ➢ **P.S, Any PC fans?! :P**



Some Interesting Plots - 8

Inference:

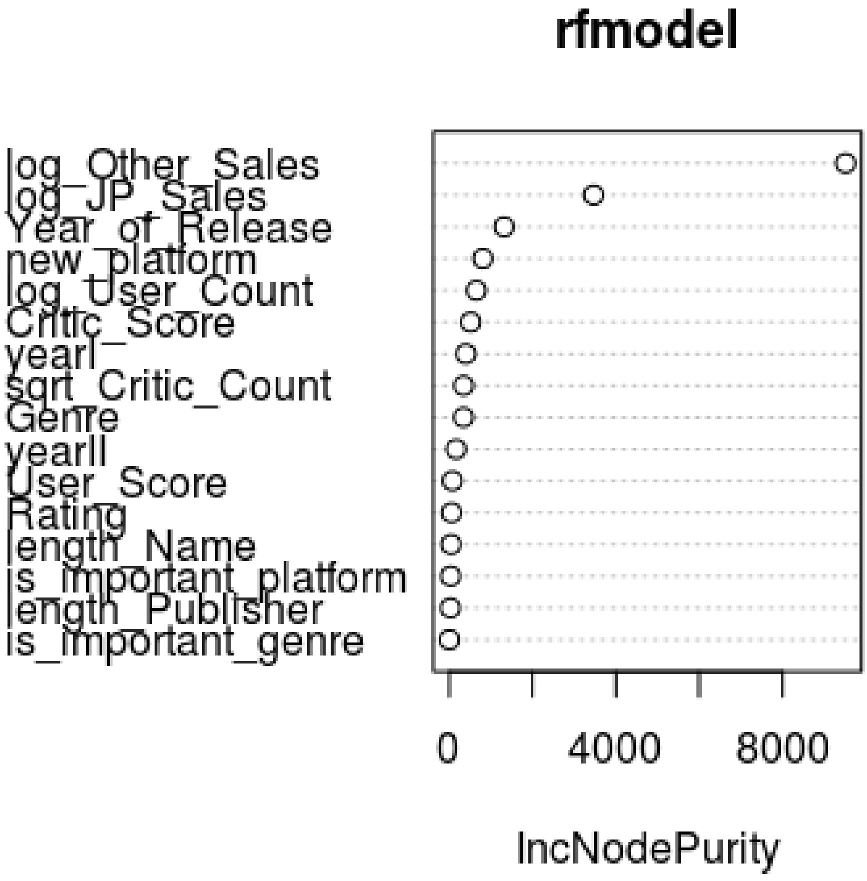
Interestingly, the length of games have increased with time!



Feature Importance – Random Forest

Inference:

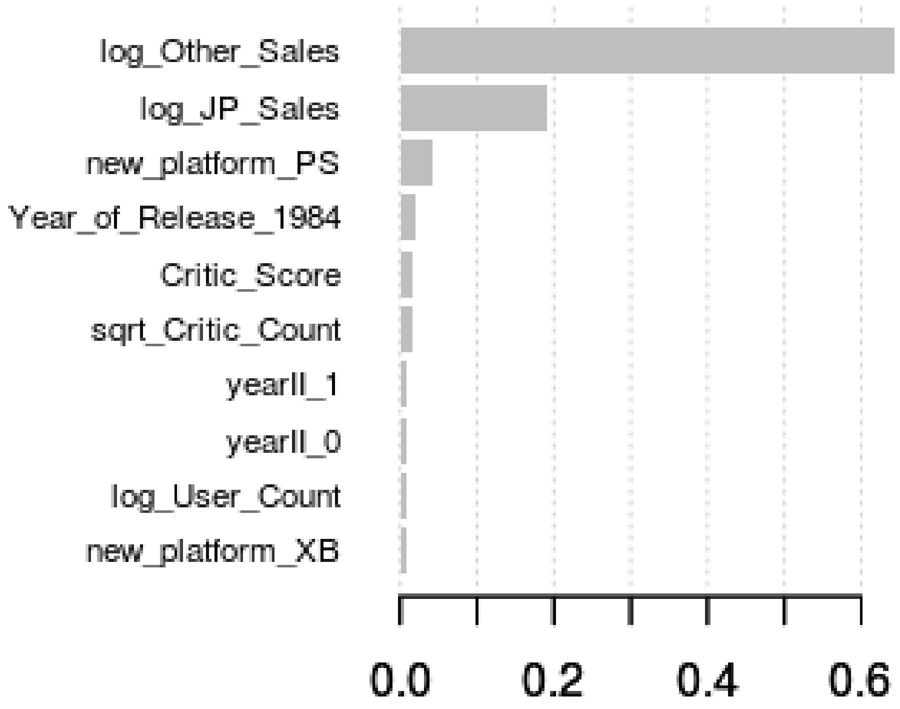
Log_Other_Sales, Log_JP_Sales, Year_of_Release are important



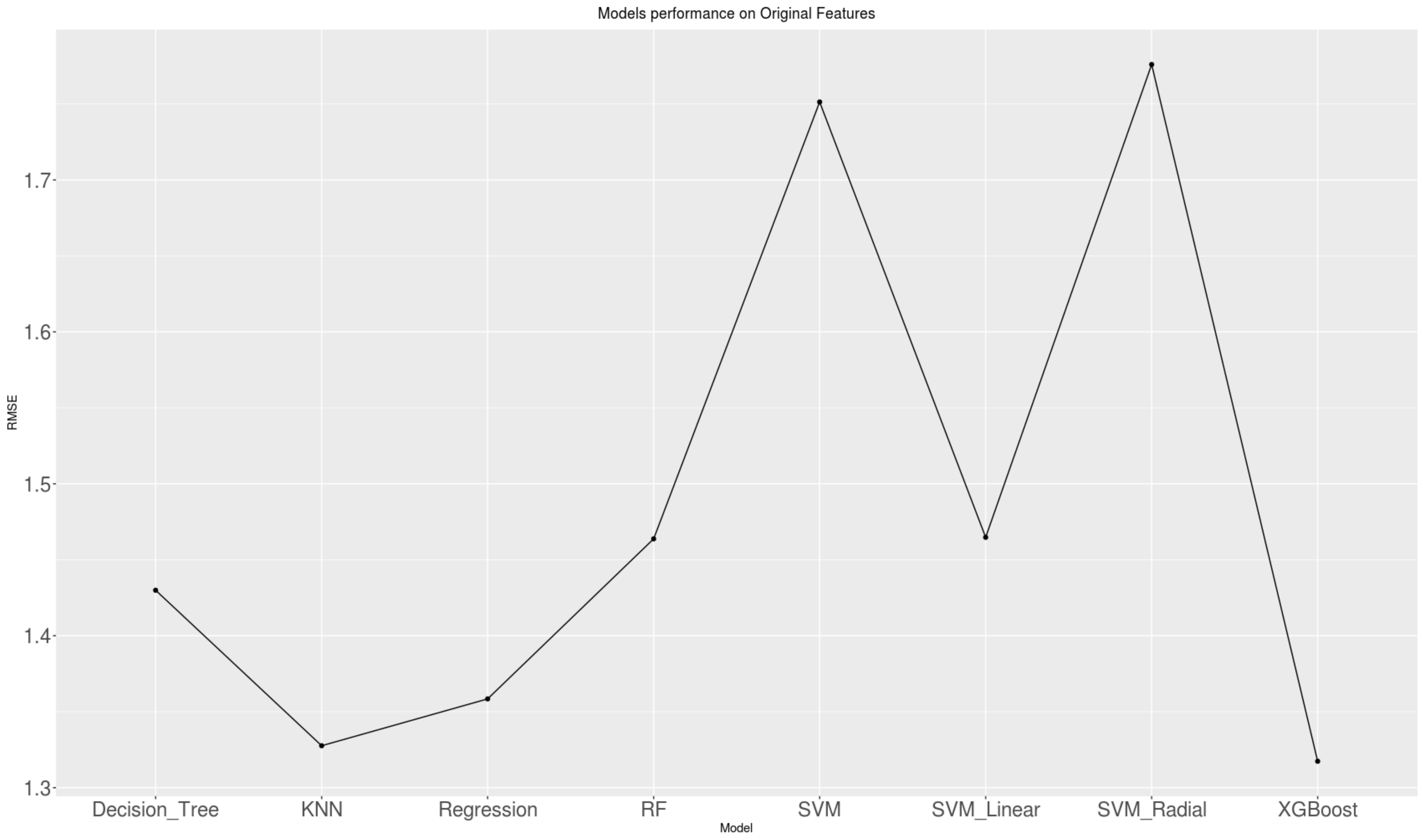
Feature Importance – XGBoost

Inference:

Log_Other_Sales, Log_JP_Sales, If Platform is PS are important features.



Models performance on Original Feature



Features from Neural Nets – Main Idea

- **Create new features from Neural Nets and run other models on them.**

Features from Neural Nets - How we do it?

- **Extracting the activation function’s output when we pass an observation.**

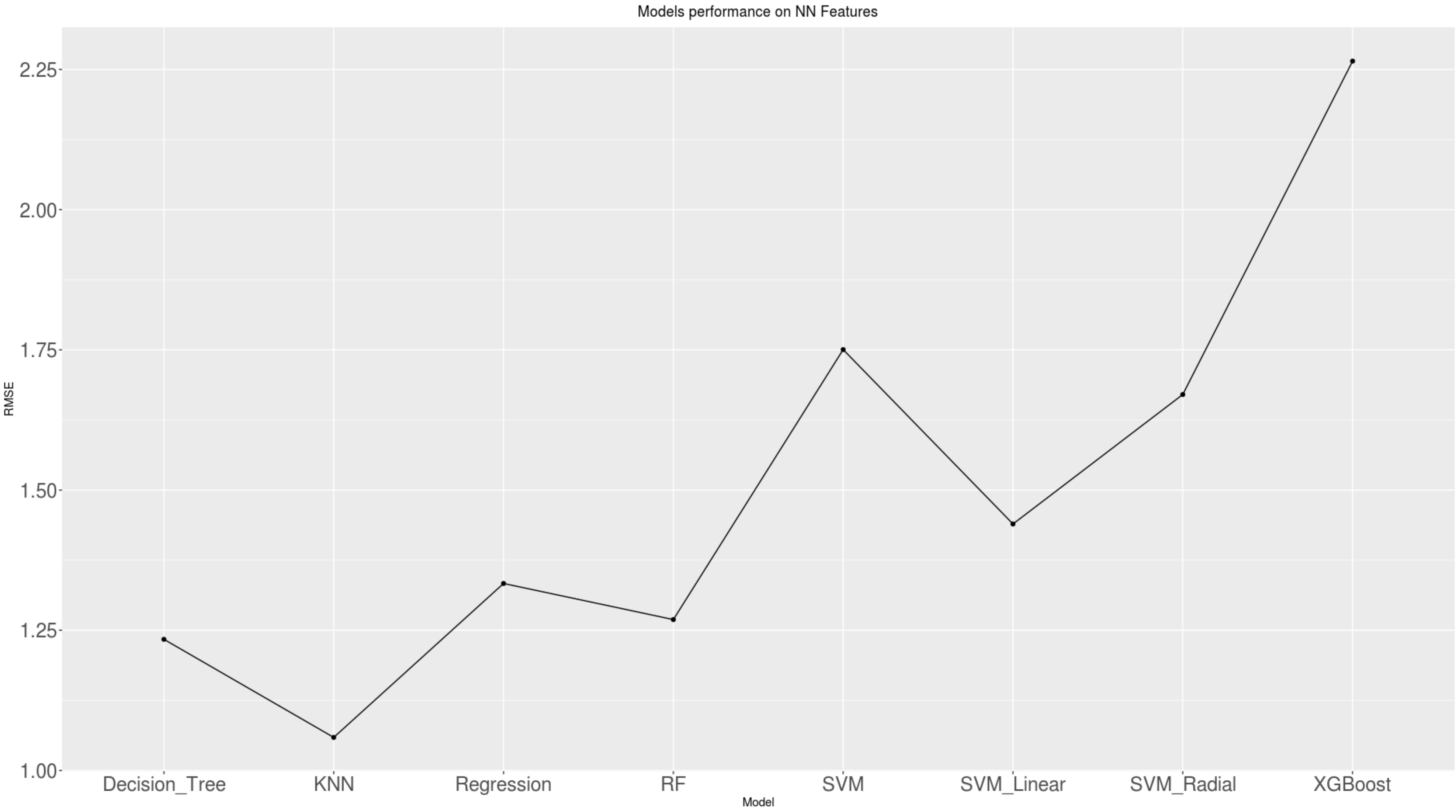
- **Neural Network:**

  ActivationFunction : **ReLUWithDropout**

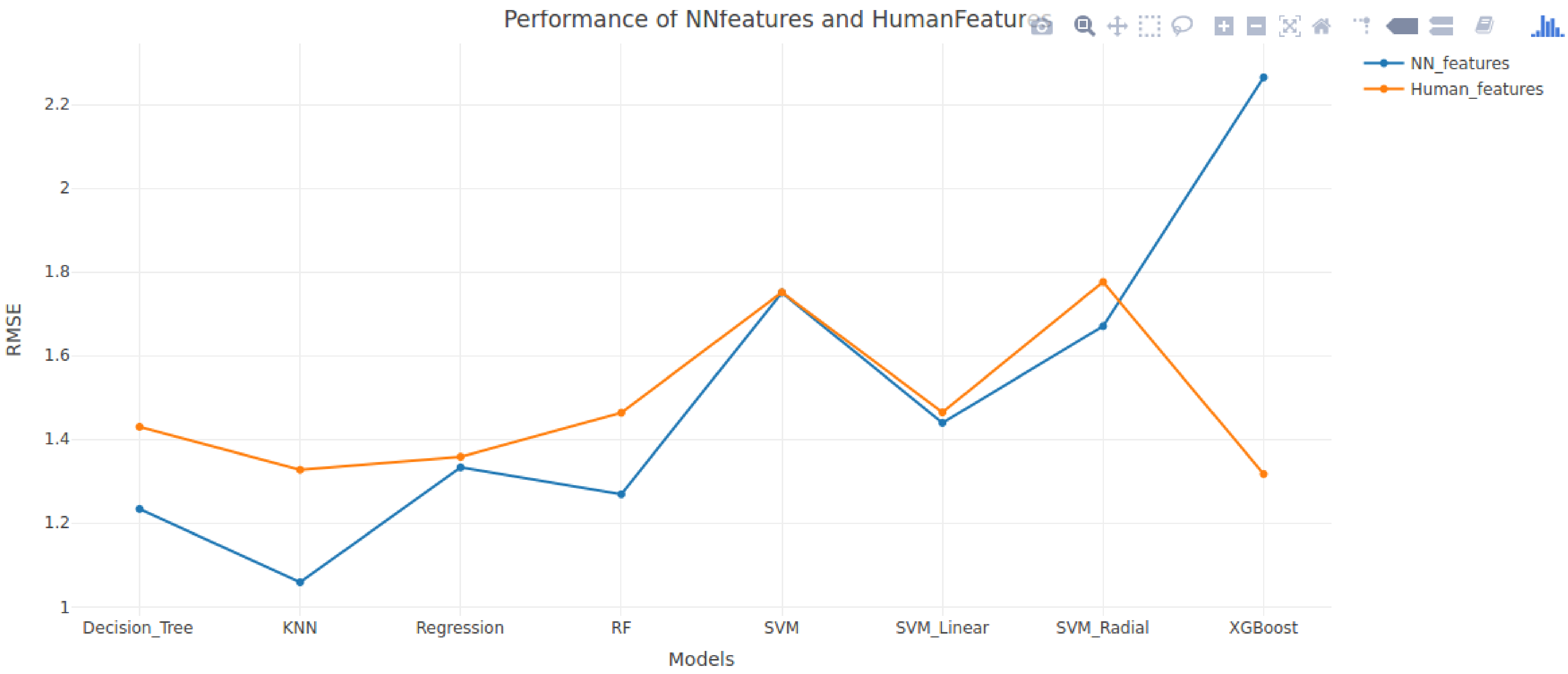
  Hidden Layers : **3 (25, 20, 10)**

  Number of Features: **55**

Models performance on NN features



FaceOff!


Performance of NNfeatures and HumanFeatures
NN_features
Human_features
RMSE
2.2
2
1.8
1.6
1.4
1.2
1
Decision_Tree
KNN
Regression
RF
SVM
SVM_Linear
SVM_Radial
XGBoost
Models

What does it tell?

- **Features from Neural Networks are clearly superior!**

Any Questions?