# Long Jumping the Sports Earnings Gender Gap

Andrew Baker
Catholic University
DA 515 – Spring 2022
abaker178@gmail.com

## Abstract

This project examines the variances between gender earnings in the world's top athletes of 2021. Using male athlete earnings data from 1990-2021, I generated an ensemble of five regression models that are used to predict the earnings of the top 10 female athletes in 2021. Contrary to typical machine learning models, the comparison of the predicted versus the actual is not to evaluate accuracy, but rather quantify the inaccuracy, ergo the inequality.

## 1 Introduction

Picture an athlete. Now try to imagine how much money that individual earned in 2021. Regardless of the demographics – age, gender, nationality, sport – the amount predicted was likely in the millions. When looking at Forbes' list of the top 10 highest-earning athletes in the world over the past 31 years, it ranges from $8.1 million (Michael Jordan in 1990) to $300 million (Floyd Mayweather in 2015) (Pandey, Forbes Highest Paid Athletes 1990-2020 2021). Interestingly though, in the past 31 years, only one female athlete ever made the prestigious list: Monica Seles in 1992. She may have made the list in subsequent years, however she was literally stabbed in the back during a tennis match in 1993 at the age of 19 and did not play again until 1995 (Editors 2019). In other words, out of the 310 total athletes in the Forbes' list, 309 of them are male.

What makes them so much different than their female counterparts? If an up-and-coming tennis player from the United States forgoes college to compete in the Olympics, turns pro at 18, and plays for three years becoming a top performing athlete, does it matter what their gender is regarding how much money they will earn? The machine learning models I built quantifies the magnitude of this discrepancy and inequality.

## 2 Disclaimers

Throughout this paper, the term ‘earnings’ refers to the sum of an athlete’s salary and endorsements. The independent testing of these two sub-categories was not taken into consideration for the models; earnings is the only target value of each model. Other features not included in the models which may have an impact on earnings include:

1. **Success**: The measure of success an athlete has in their sport (e.g. number of championships or titles won or ranking in relation to all athletes in the sport)
2. **Social Media Presence**: How active/followed an athlete is on social media may have an impact on their endorsements and therefore their total earnings
3. **Country of Residence:** The current country in which the athlete is residing

There is also an assumption that the top 10 female athletes are working, performing, competing, and succeeding at the same level as the top 10 male athletes.

## 3 Data

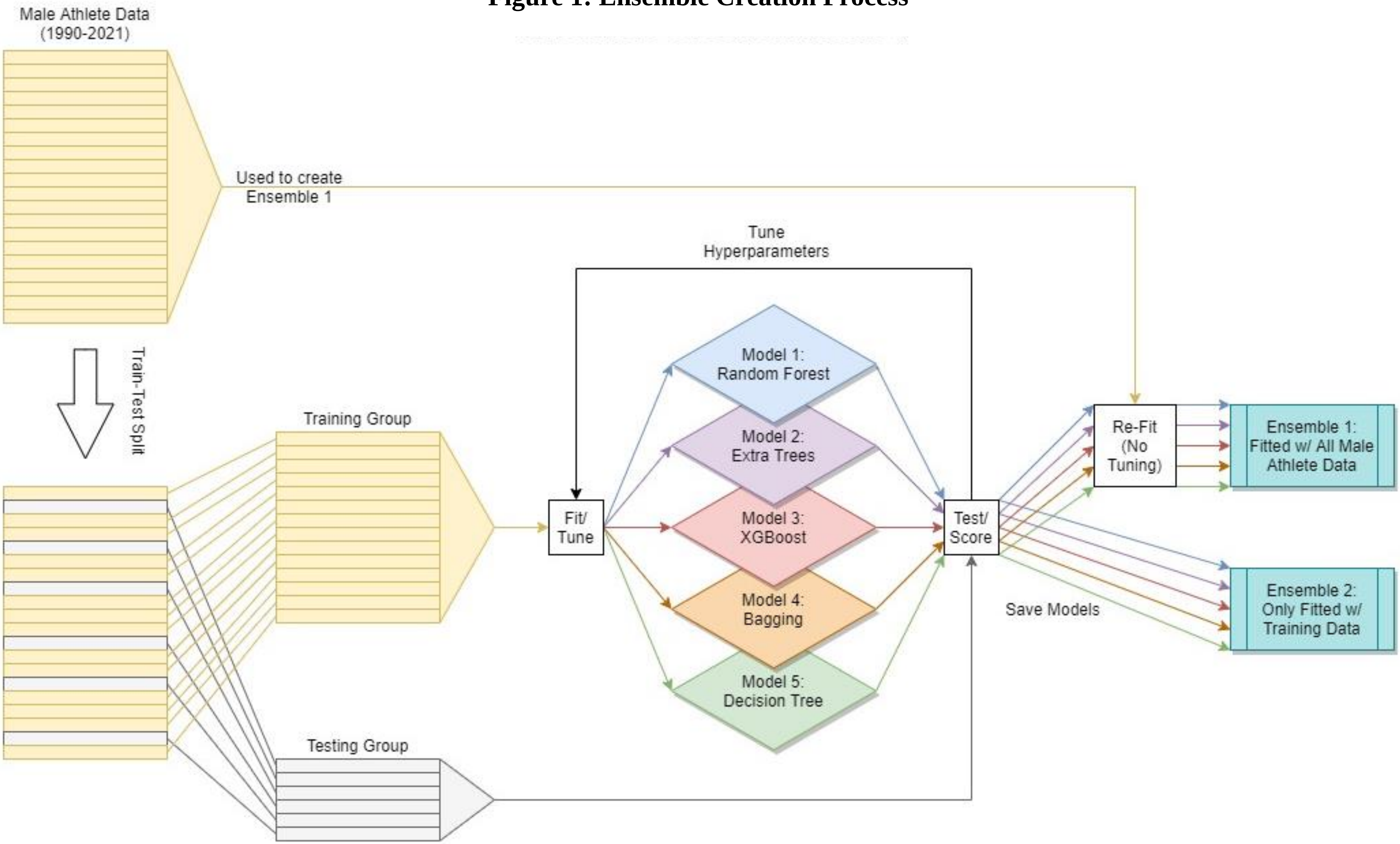
The primary source of data used for building the models comes from Forbes’ list of the top 10 highest-earning athletes from 1990-2020 accessed via Kaggle (Pandey 2021). Earnings data for 2021 was extracted from Knight’s article on Forbes (Knight 2021). Additionally, I collected the data for the top 10 highest-earning female athletes in 2021 from Knight on Forbes (Knight, The Highest-Paid Female Athletes Score A Record $167 Million 2022). Lastly, Forbes provides the top 50 highest-earning athletes in 2021 (Knight and Birnbaum, 2021 Highest-Paid Athletes 2021). Supplemental feature data, not included in this dataset, was manually scraped by researching each athlete.

The primary dataset (top 10 from 1990-2021) was split into 75% training and 25% validation/testing sets to build each model. Once the hyperparameters were tuned (see Table 1 for hyperparameter selection by model), the 2021 top 10 female-only data is used to obtain their predicted earnings: a quantitative continuous variable.

**Table 1: Hyperparameter Selection for Regressions**

| Model | n Estimators | Criterion | Splitter | Max Depth | Minimum Samples Split | Minimum Samples Leaf |
|---|---|---|---|---|---|---|
| Random Forest | 100 | mse | - | 11 | 2 | 1 |
| Extra Trees | 100 | mse | - | 17 | 2 | 1 |
| XGBoost | - | - | - | 10 | - | - |
| Bagging | 15 | - | - | - | - | - |
| Decision Tree | - | friedman_mse | Best | 16 | 2 | 1 |

## Figure 1: Ensemble Creation Process

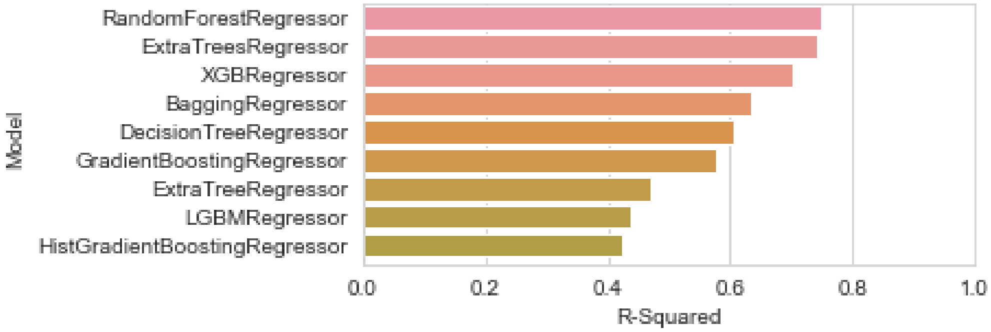


# 4 Models

Two ensembles are created both of which are comprised of the same five models tuned identically. The difference is the data which is fit to each model. Each model in Ensemble 1 is fit with all the original data (the top 10 athletes from 1990-2021). Ensemble 2 models are only fit with the 75% subset used to train the models (see Figure 1 for full ensemble creation process). The purpose of this is to ensure I could fit the models with all the original data and not drastically change the predicted outcomes while still fitting the models with as much applicable data as possible. Once Ensemble 1 (all data) is confirmed to have a minimal variance from Ensemble 2 (training data only), Ensemble 1 is used as the only testing ensemble.

LazyPredict is used to find the top five applicable models. The top results with the best $r^2$ scores were all regressions (Figure 2). Each ensemble is built using only the top five: Random Forest, Extra Trees, XGBoost, Bagging, and Decision Tree.

**Figure 2: LazyPredict Results**



## 4.1 Random Forest Regression

According to the Lazy Predict results, the model with the greatest $r^2$ value of 0.75 is Random Forest Regression. This model type is an ensemble comprised of many Decision Trees (see section 4.5) all run sequentially, and the results averaged to yield a more consistently accurate model than the individual trees. A Random Forest ensemble subsets the training data for each tree, as well as selects the best feature on which to split a node. After tuning, the $r^2$ value is increased to 0.97 (Figure 3).

## 4.2 Extra Trees Regression

The Extra Trees Regression comes in second in the LazyPredict $r^2$ values with a value of 0.74. Extra Trees (short for Extremely Randomized Trees) is very similar to Random Forest with a few slight differences. Extra Trees uses the entirety of the provided training data for each tree in the ensemble and randomly chooses the feature on which to split a node. As with Random Forest Regressions, the results of all the Decision Trees are averaged together and returned as the output for the Extra Trees model. After tuning, the $r^2$ value increased to 0.999 rounded to 1.0 (Figure 4).

## 4.3 XGBoost Regression

XGBoost comes third in the LazyPredict list with an $r^2$ value of 0.70 and is the only gradient tree model in the ensemble. Gradient Boosting combines an array of simpler models which are weaker in comparison, but when combined, increase the target accuracy (AWS.com 2022). In the case of regression, XGBoost uses regression trees (see section 4.5) and "a gradient descent algorithm to minimize the loss when adding new models" or trees (AWS.com 2022). After tuning, the $r^2$ value improved to 0.999 rounded to 1.0 (Figure 5).

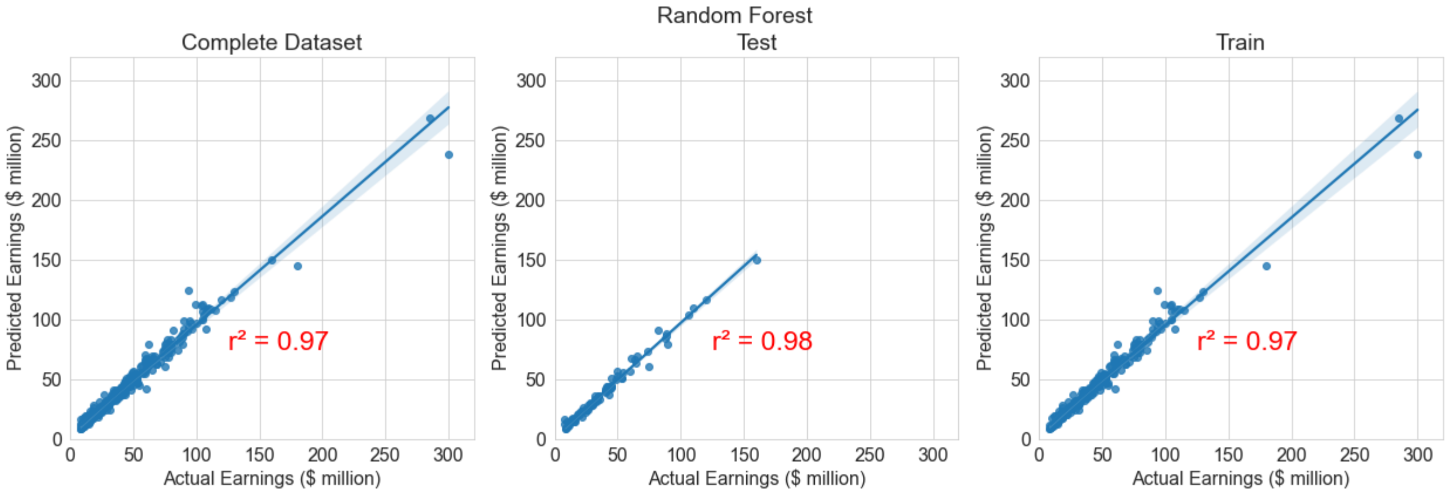
## 4.4 Bagging Regression

A Bagging Regression model ($r^2 = 0.63$) is a simple Linear Regression ensemble that considers all points of the data simultaneously. Bagging partitions the data into "bags", runs the Linear Regression algorithm on each, and averages the results which is then returned as the result of the ensemble (Amrit 2020). This method of splitting the data lowers variance while not raising bias (Amrit 2020). As compared to the standard Linear Regression model which scored an $r^2$ value of -2.6e24, the bagged method allows use of a model that does not inherently work with this dataset. After tuning, the $r^2$ value improved to 0.96 (Figure 6).

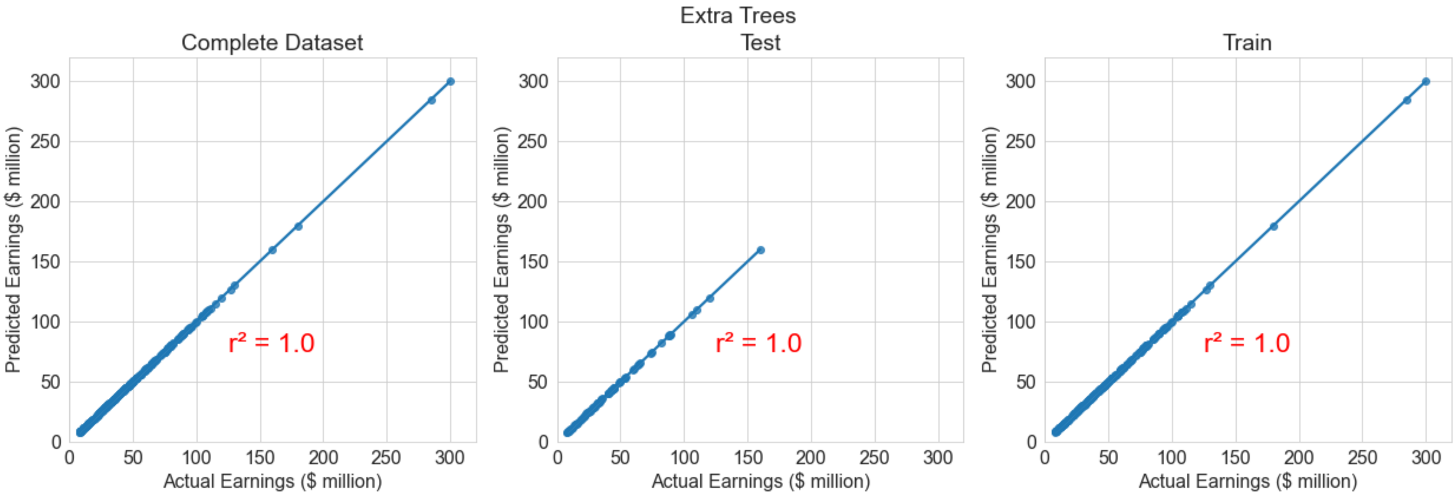
## 4.5 Decision Tree Regression

The Decision Tree Regression model is the basis for two other models and scores just above my cutoff of 0.60 with a LazyPredict $r^2$ value of 0.61. A Decision Tree considers all data provided and creates a split point dividing it into two or more nodes. With any Decision Tree, a goal is to find the right depth for the data. In terms of a regression, once the best depth for the data is reached, all target values (earnings in this case) in each node are averaged together. Any testing feature data will follow the splits down into one of the final groups and its predicted value will be that average. After tuning, the $r^2$ value increased to 0.999 rounded to 1.0 (Figure 7).

## Figure 3: Random Forest Regression Results



## Figure 4: Extra Trees Regression Results



## Figure 5: XGBoost Regression Results

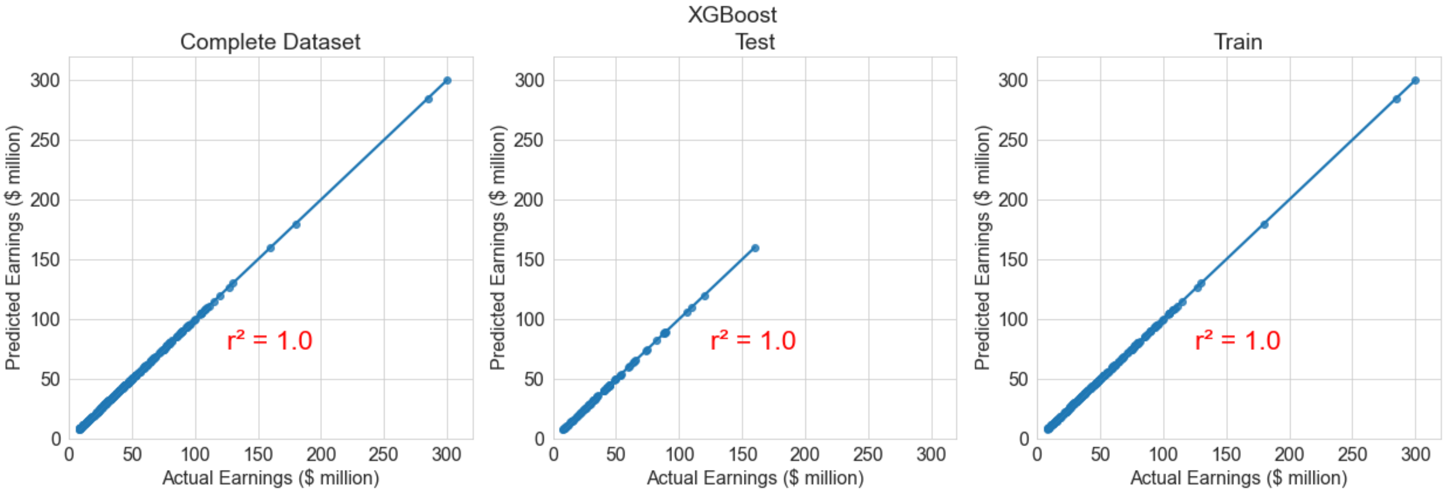


**Figure 6: Bagging Regression Results**

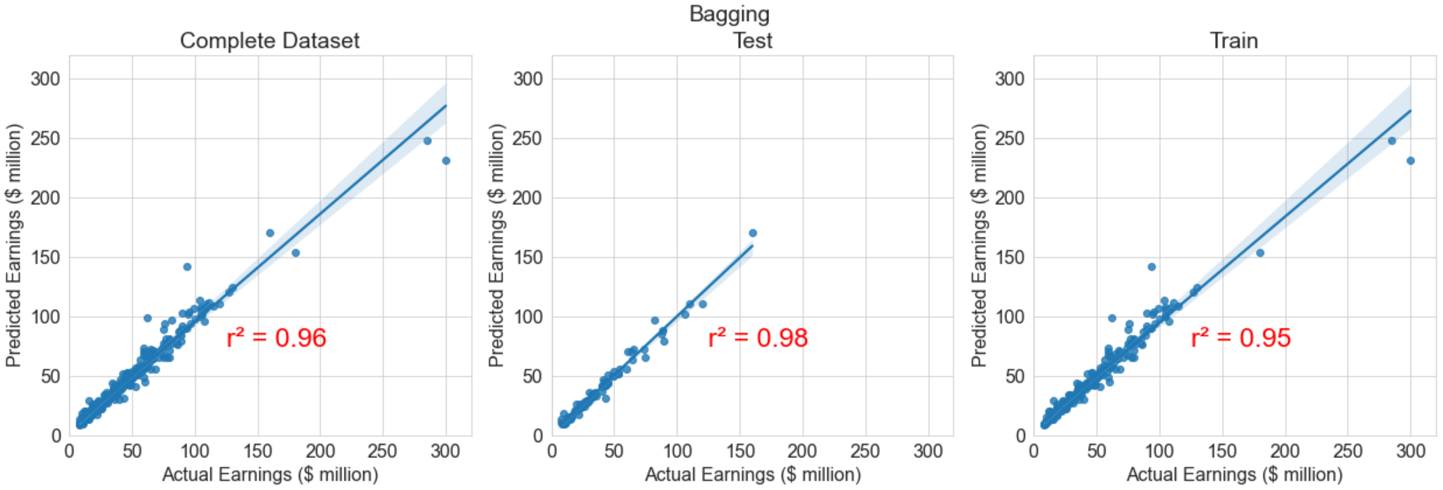
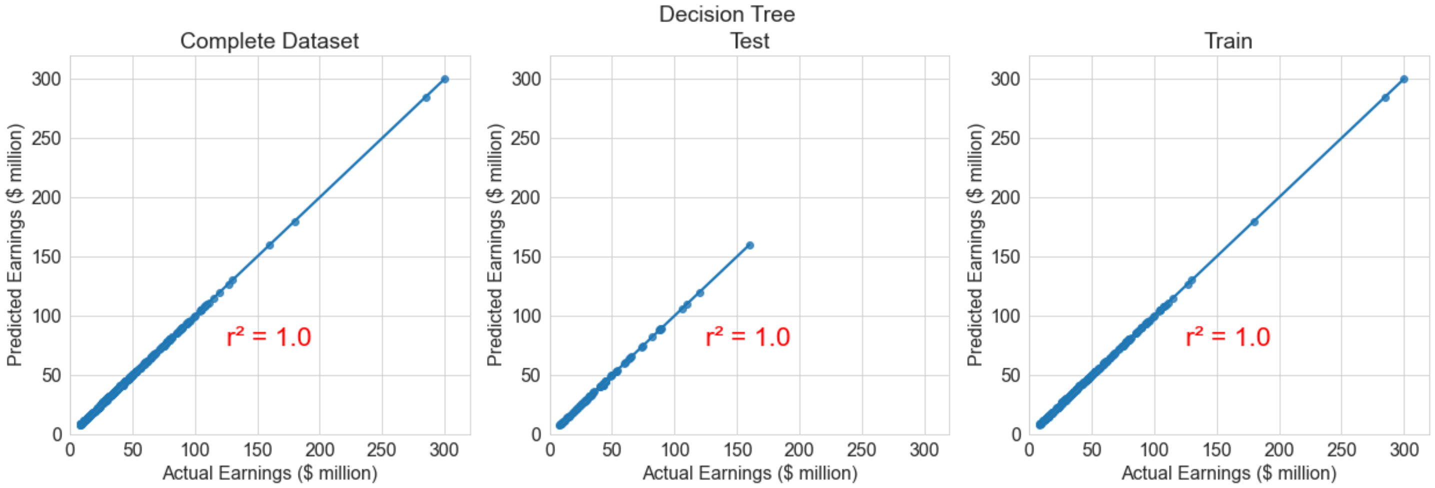


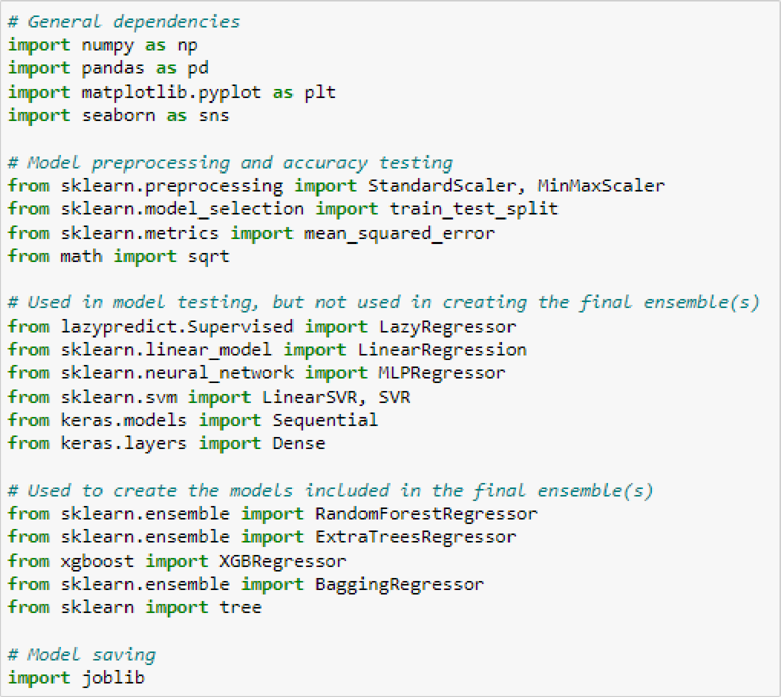
**Figure 7: Decision Tree Regression Results**



## 5 Implementation

All code was written in Python 3.8 using the Sci-Kit Learn’s submodules such as preprocessing, model selection, and ensemble. The models are out-of-the-box with only hyperparameter tuning being performed on each. The Trees and Bagging are from Sci-Kit Learn. XGBoost was developed by Distributed (Deep) Machine Learning Community (DMLC) and made available through Apache License 2.0. All scripts were performed on a local CPU. A complete list of modules used can be found in Figure 8.

## Figure 8: Complete Module List



```
# General dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Model preprocessing and accuracy testing
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt

# Used in model testing, but not used in creating the final ensemble(s)
from lazypredict.Supervised import LazyRegressor
from sklearn.linear_model import LinearRegression
from sklearn.neural_network import MLPRegressor
from sklearn.svm import LinearSVR, SVR
from keras.models import Sequential
from keras.layers import Dense

# Used to create the models included in the final ensemble(s)
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import ExtraTreesRegressor
from xgboost import XGBRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn import tree

# Model saving
import joblib
```

# 6 Results

There are two aspects to the results for this project:

1. The accuracy of each model during the tuning process; and
2. The expected *inaccuracy* of the ensemble.

## 6.1 Model Accuracy (Build & Tuning)

We must build accurate models that can predict earnings based on a given set of features (not including gender) before we can show that all things being equal, a female athlete is not earning the same as their male athlete counterpart. The features included in the models in alphabetical order are as follows:

1. **Age**
2. **Birth Year**
3. **Earned College Degree**: Whether the athlete earned a college degree *prior* to the earnings report. Note: an honorary degree does not count
4. **Nationality**: The athlete's birth country
5. **Olympian**: Whether the athlete is an Olympian
6. **Sport**
7. **Went Pro Age**: Age at which the athlete became a professional
8. **Year**: The year of the earnings report
9. **Year Went Pro**: The year the athlete became a professional
10. **Years Pro**: The number of years as a professional at the time of the earnings report

The training ratio for building each model is 70%; 30% for testing. After tuning the hyperparameters for each model and scoring its predictions for the entire dataset, the range of $r^2$ values is 0.96 to 1.0. The $r^2$ values for the testing set ranged from 0.98 to 1.0 (Figures 3-7). With such high coefficients of determination, I was comfortable with each individual model's ability to predict earnings based off this feature set.

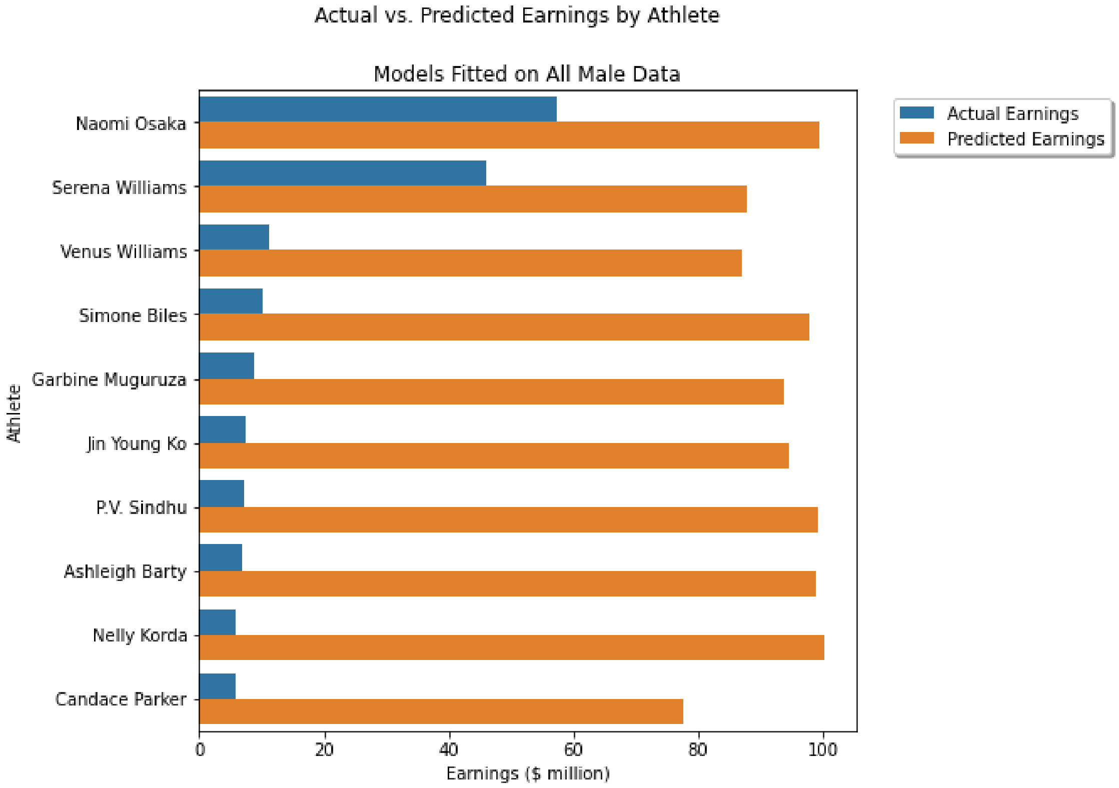
### 6.2 Testing for Inaccuracy; Highlighting the Inequality

Normally, with the models trained and the ensemble ready to test another dataset, the desired result to support a hypothesis would be to see similarly high accuracy values. My hypothesis, however, is the predictions will *not* be close to the actual target values given the feature data and thus will highlight the inequality in what female athletes earn versus what a highly accurate ensemble predicts they earn. I ask the reader to suspend the notion that extreme inaccuracy in applying new feature data to an ensemble is a result of an inaccurate ensemble, and instead, focus on the severe discrepancy given a very important missing feature: gender.

The top 10 highest-earning male athletes averaged $105.2 million in 2021. According to the ensemble, the average predicted earnings for the top 10 female athletes in 2021 is $95.68 million. The *actual* average earnings of those same female athletes is $16.66 million; a difference of $79 million.

A complete breakdown of each female athlete's actual earnings versus the predicted earnings can be seen in Figure 9.

## Figure 9: 2021 Top 10 Highest-Earning Female Athletes' Actual and Predicted Earnings



## 7 Conclusions

Even if the unavailable features mentioned in Disclaimers had zero impact on the building of the models, we would still not be comparing apples to apples. In 2021, the top 48 highest-earning male athletes made an average of 71% of their earnings from salary, leaving 29% from endorsements. The top 10 female athletes only made an average of 19% from salary, leaving 81% earned from endorsements (Figure 10). Despite the efforts to build an environment where we could say, with all things being equal, this is what the top female athletes should be making, we still end up with the fact that male and female athletes *earn* their money differently. Endorsements drove the earnings of female athletes in 2021 while their male counterparts are getting paid to play their sport, not be a spokesperson. The $r^2$ value of 0.002 when comparing the actual and expected earnings of the top female athletes supports this finding. Hidden from the models and ensemble, gender is the only characteristic changed from our original model/ensemble testing, and yet the chaos introduced by this adjustment is remarkable. A factor of this inconsistency could be the nature of endorsement amounts and how they do not always rely on the athlete's performance, but rather the public's view of the athlete or even just the sport they play (Rascher, Eddy and OSKR 2017).

A closer examination of the top female athletes reveals an interesting consistency. Each of these athletes are Olympians in a Summer Games sport (tennis, badminton, golf, and basketball). Given that the 2020 Summer Olympic Games were rescheduled to 2021 due to the COVID-19 pandemic (Talmazan 2020), the endorsement deals that may have derived from the games would have been included in the 2021 earnings report. This begs the question, "would these specific female athletes have made the 2021 top 10 list had the games taken place in 2020 as originally scheduled?" I speculate that this list would have looked very different.

Furthermore, given that 81% of the top 10 female athlete earnings came from endorsements, the inclusion of the Summer Games in the earnings report – and the list being entirely comprised of Summer Games Olympians – depicts the disparity between the popularity of the Olympics versus women's professional sports leagues and is not a reflection of their prowess in their individual sport. This suggests that we, as a society, only want to watch and support women athletes when they are in the Olympics. For transparency, only 6 of the top 10 males are Olympians and they are all summer sport athletes as well.

Profitability of an athlete depends on many factors, including viewership. Until women's sporting events are advertised, televised, and attended more, the primary reason for the salary gap will remain. Increasing these three aspects will create a cascading profitability increase of the sport, the teams, and the players.

## 8 Future

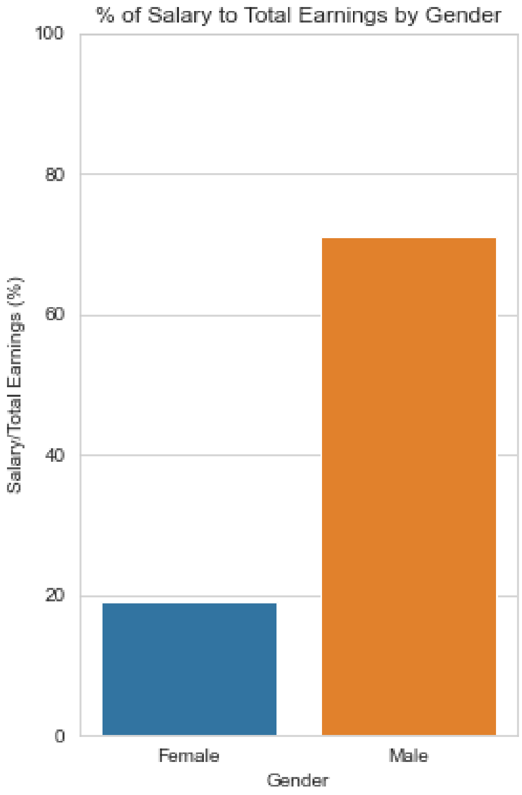
Given the fact that Forbes' 2021 top 10 female earners were all Olympians in summer sports and females typically earn most of their money through endorsements rather than salary, I will be rerunning this experiment each year Forbes releases these metrics. I am especially interested in 2022's results as it will contain the 2022 Winter Olympics in

Beijing which took place between February 4th and February 20th (olympics.com 2022). My initial hypothesis for the results of Forbes' top 10 female athletes will be that they are all Olympians from winter sports.

2023 will be even more interesting to investigate as it will be a year in which no Olympics take place and should give a very transparent view of the true inequality between male and female athletes. What will each gender's athletes look like – earnings-wise – when the world and the global markets are not hyper-focused on Olympians? I imagine the gender-gap will be even greater than it is in 2021 and what I predict it will be in 2022.

From the perspective of improving the machine learning component, the inclusion of more features (such as those listed in Disclaimers) may allow other model types besides trees to become just as useful to the ensemble. This 'widening' of the feature set should permit use of a Support Vector Regression that relies on many features for its accuracy. Over time, the number of data points will gradually increase, however, at a rate of 10 athletes per year, it would take a millennium to get to 10,000 data points. Therefore, it would be more realistic to broaden the scope to the top 50 or top 100 highest-earning athletes per year to have enough data points to use a Neural Network. As well, focusing solely on salary instead of total earnings may yield a better Linear Regression model as salary *should* be more indicative than endorsements as it relates to an athlete's performance. Lastly, the inverse to the premise of this project would be interesting – collect the top 10 highest-earning female athlete data from the past 30 years and repeat the process to see what the model predicts the male athletes should earn.

## Figure 10: Percent of Salary to Total Earnings by Gender



## 9 References

Amrit, Akshay. 2020. *Bagging on Low Variance Models.* October 12. Accessed April 29, 2022. https://towardsdatascience.com/bagging-on-low-variance-models-38d3c70259db#:~:text=The%20bagging%20technique%20creates%20multiple,and%20coefficients%20of%20every%20model.

AWS.com. 2022. *How XGBoost Works.* Accessed May 1, 2022. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html.

Editors, History.com. 2019. *Tennis star Monica Seles stabbed.* A&E Television Networks. July 28. Accessed April 30, 2022. https://www.history.com/this-day-in-history/tennis-star-monica-seles-stabbed.

Knight, Brett. 2022. *The Highest-Paid Female Athletes Score A Record $167 Million.* January 12. Accessed April 2022. https://www.forbes.com/sites/brettknight/2022/01/13/the-highest-paid-female-athletes-score-a-record-167-million/?sh=455f680378cc.

—. 2021. *The World's 10 Highest-Paid Athletes.* Accessed April 2021. https://www.forbes.com/sites/brettknight/2021/05/12/the-worlds-10-highest-paid-athletes-conor-mcgregor-leads-a-group-of-sports-stars-unfazed-by-the-pandemic/?sh=2aa92d4926f4.

Knight, Brett, and Justin Birnbaum. 2021. *2021 Highest-Paid Athletes.* Accessed April 2022. https://www.forbes.com/athletes/.

olympics.com. 2022. *Beijing 2022.* Accessed April 30, 2022. https://olympics.com/en/olympic-games/beijing-2022.

Pandey, Parul. 2021. *Forbes Highest Paid Athletes 1990-2020.* Accessed March 2022. https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019.

—. 2021. *Forbes Highest Paid Athletes 1990-2020.* Accessed March 2022. https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019.

Rascher, Daniel, Terry Eddy, and LLC OSKR. 2017. "What Drives Endorsement Earnings for Superstar Athletes?" *Sport Management* 9. https://repository.usfca.edu/sm/9.

Talmazan, Yuliya. 2020. *New dates announced for Tokyo 2020 Olympics postponed over coronavirus concerns.* March 30. Accessed May 1, 2022. https://www.nbcnews.com/news/world/new-dates-announced-tokyo-2020-olympics-postponed-over-coronavirus-concerns-n1171871.