Unit of Analysis for the Model

The unit of analysis that we used for the model was predicting whether teams in the NCAA March Madness Tournament would win or lose their games. The purpose of creating this model is to better understand whether teams will win or lose their games but also use the model to predict how teams in this year's March Madness Tournament will perform during each round.

Data Used to Train and Test the Model

The data used to train and test the model was web scraped from the ESPN site using Python. The timeframe of the data pulled was from November 6th to February 10th. We used this timeframe because it would give us a range of how the teams have performed over a couple of months. We used the following data columns to test and train the model: Team, Opponent, Points (Home Team), Opponent Points, Home Point Differential, Win (Did the team win?), Row Count (counting how many games the team has played from November 6th to February 10th), Average Point Differential, and Total Wins.

How the Data was Cleaned and Aggregated

We obtained our dataset from ESPN's game data through a process of screen scraping. Utilizing code developed by an individual who had achieved the highest data retrieval success, we were able to acquire comprehensive data. With Python, we then generated a rolling average. However, we encountered a challenge in calculating the rolling average for a team's first three games, which inadvertently included the last team's game data. To address this, we implemented an IF statement in Excel, resolving the issue. Additionally, we employed Excel to compute the running total of team wins.

Type of Analysis Conducted

We used logistic regression, decision tree analysis, and random forest analysis in the model. Here are the reasons for using these forms of analysis:

Logistic regression: It draws the most straightforward conclusions from the data.

Decision Tree Analysis: It incorporates non linear relationships that can be easily understood when mapped out and visualized.

Random Forest Analysis: Depending on potential limitations such as overfitting and outliers this can be better than a decision tree.

What Factors are Included in the Analysis

In our analysis of predicting the winner of the NCAA March Madness tournament we used logistic regression, decision tree analysis, and random forest analysis. The factors included in the analysis were the team's performance metrics, games won or lost, average point differential, and total wins. These factors were included in the analysis to help us predict the winners of upcoming games.

Results of the Analysis

Classification accuracy

Scores

Model	AUC	CA
Logistic Regression	1.000	1.000
Random Forest	1.000	1.000
Tree	1.000	1.000

Every model received a classification accuracy (CA) of 1.0. In other words, every data point has been classified correctly according to the model's predictions. This indicates that the model is performing incredibly well.

Confusion Matrix

Confusion matrix for Logistic Regression (showing number of instances)

		Predicted		
		0	**1**	**∑**
Actual	**0**	4467	0	**4,467**
	1	2	4469	**4,471**
	∑	**4,469**	**4,469**	**8,938**

The confusion matrix for this model shows 4,467 correct "loss" predictions and 4,469 correct "yes" predictions with only 2 incorrectly predicted losses which were actually wins.

Random Forest

Confusion matrix for Random Forest (showing number of instances)

		Predicted		
		0	**1**	**∑**
Actual	**0**	4467	0	**4,467**
	1	4	4467	**4,471**
	∑	**4,471**	**4,467**	**8,938**

The confusion matrix for this model shows 4,467 correct "loss" predictions and 4467 correct "yes" predictions with only 4 incorrectly predicted losses which were actually wins.

Decision Tree

Confusion matrix for Tree (showing number of instances)

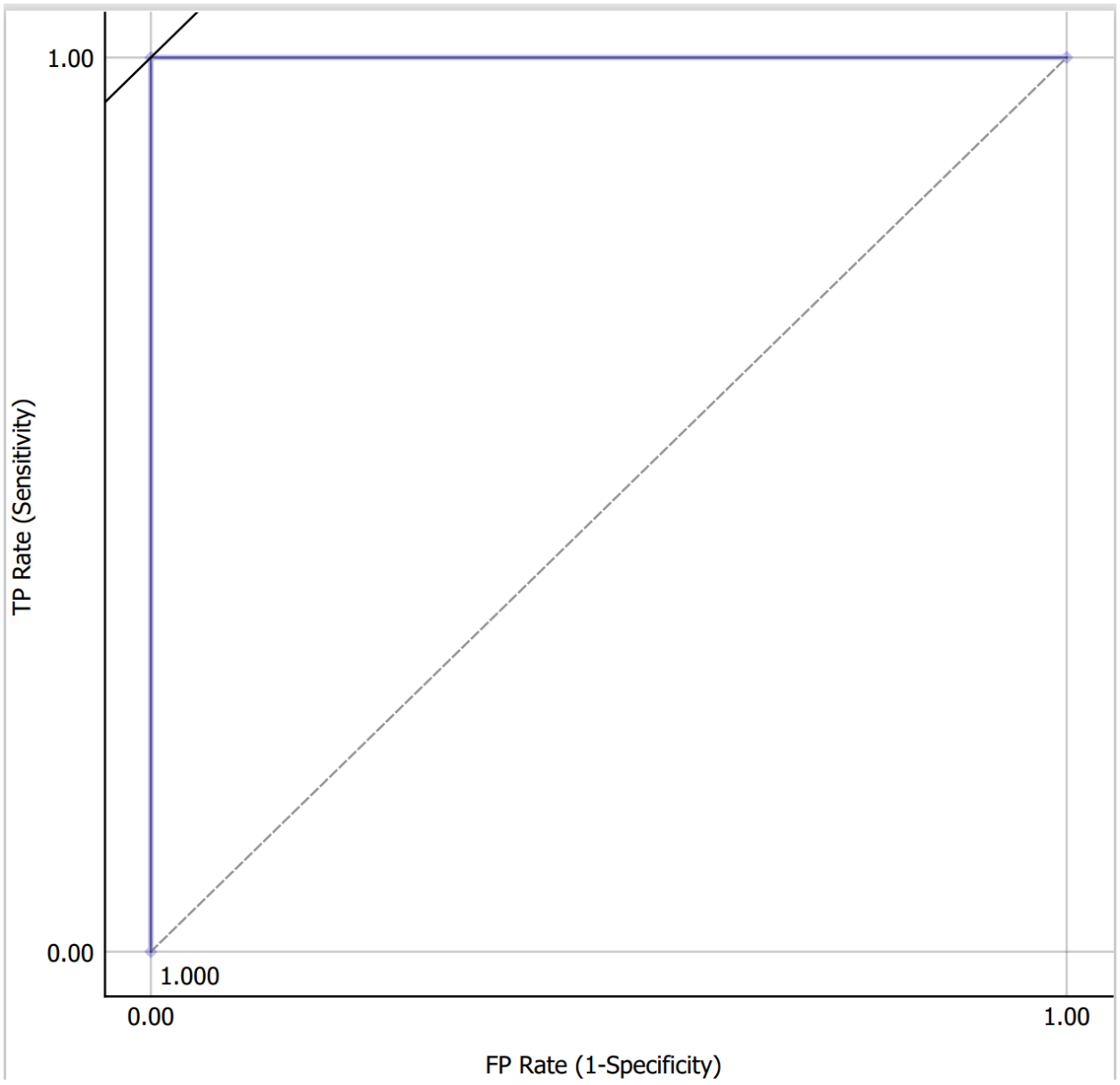
		Predicted		
		0	**1**	**∑**
Actual	**0**	4467	0	**4,467**
	1	0	4471	**4,471**
	∑	**4,467**	**4,471**	**8,938**

The confusion matrix for this model shows 4,467 correct “loss” predictions and 4,471 correct “yes” predictions with only 2 incorrectly predicted losses which were actually wins.

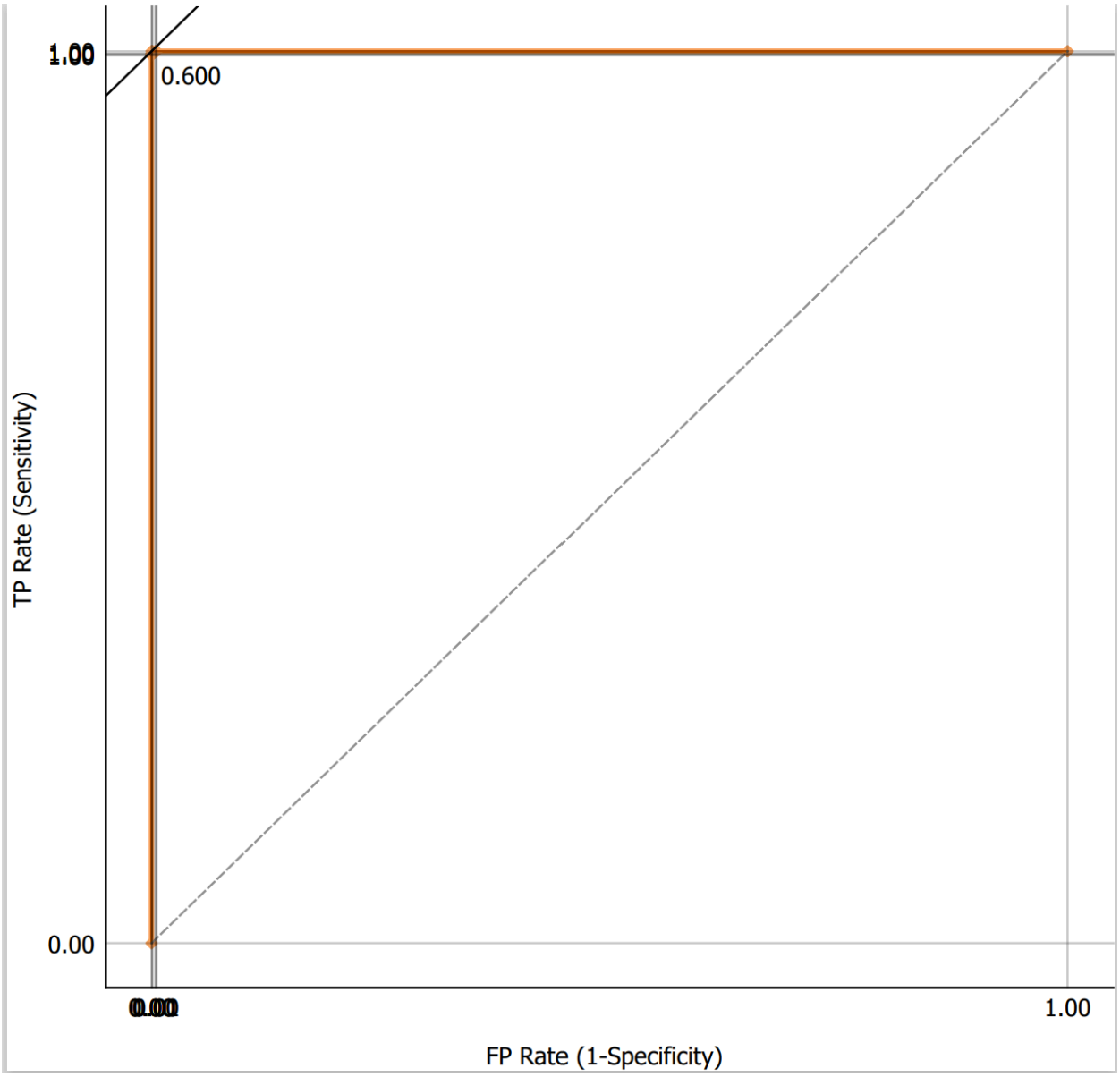
ROC Curve

The ROC curve illustrates the trade-off between true positive rate and false positive rate across different classification thresholds, with a higher area under the curve (AUC) indicating better model performance. A curve closer to the top-left corner signifies superior discrimination ability of the model. Thus, the following models show models with “superior” discrimination ability.

Decision Tree



Random Forest



Logistic Regression