Cs 118 Data Science

Final Project

Harry Myers
12/4/2024

1. The data set: atp_matches_2023.csv

- I found this data set on github.
- Has data on every professional tennis match since 1968.
- Made by passionate tennis fans and engineers.
- I chose data from 2023

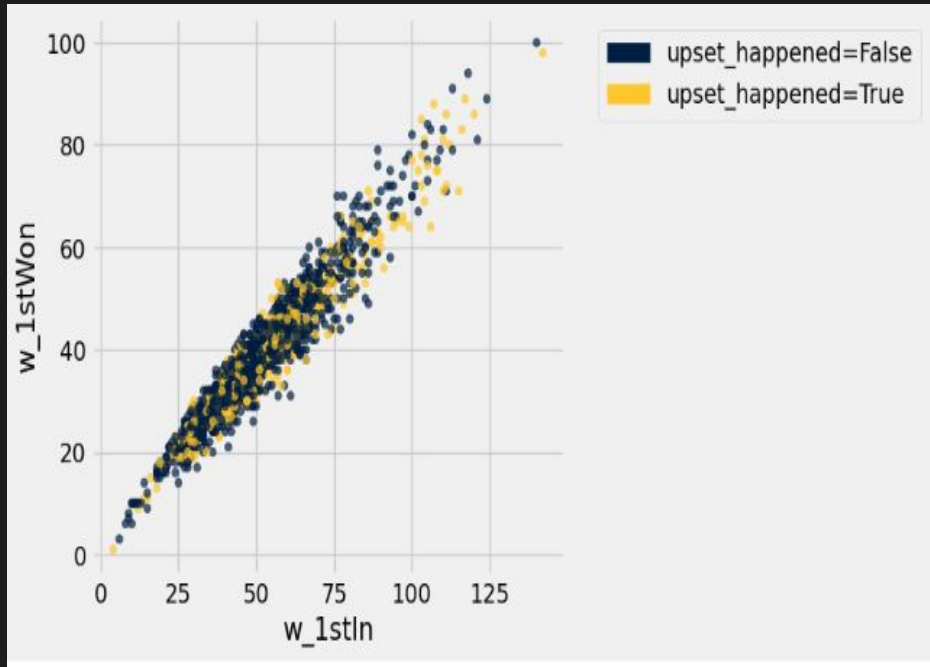
2. Regression

- Winners # first serves in vs. Winners # first service points won
- Confidence interval was 96%
- Value of x was 75 first serves in
- 99% CI: lower 54.7, upper 55.4
- Biases such as playing surface, importance of tournament

Q: How many first service points were won by a winner of an atp tour match if he made 75 first serves?

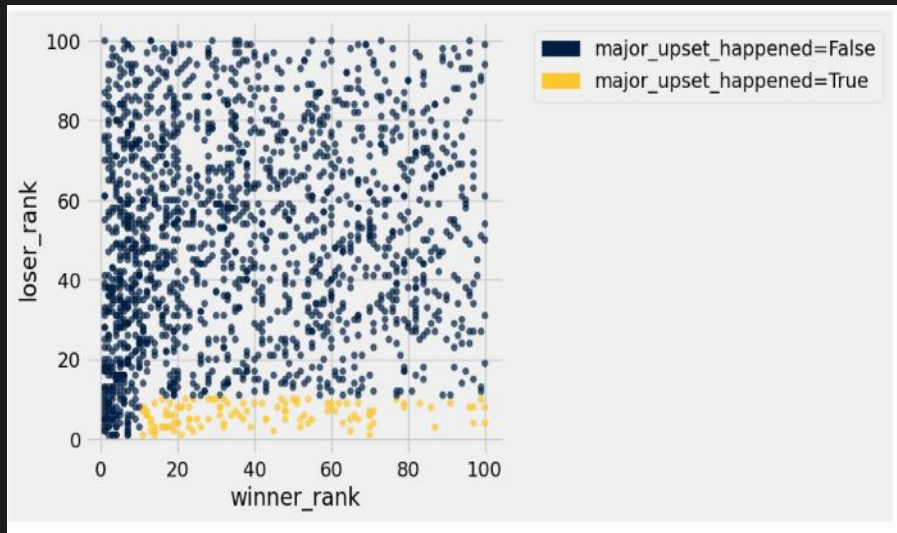
A: 55.17

Meaning around 73% of first service point were won. Pretty Good!



3. Classifier

- Classify a major upset: player outside the top ten beats a top ten player, happens ~4.5% of the time
- Features: winner rank, loser rank, winner age, loser age, match duration, winner ace count, loser ace count
- >95% accuracy, classified 715/750 correctly
- Bias: Imbalance in data set due to rarity of a major upset



4. Improved Classifier

- Old features: winner rank, loser rank, winner age, loser age, match duration, winner ace count, loser ace count
- New features: winner rank, loser rank, winner ace count, winner double fault count, winner first service count
- Changed k from 3 to 7: more neighbors might yield higher accuracy
- >97% accuracy, classified 730/750 correctly

Conclusion

- There are interesting relationships between various match statistics, such as first serves in and first serves won.
- Realized the importance of selecting the right features and values of k in K-NN which can impact the model's accuracy.
- Importance of avoiding overfitting to ensure that the model performs well on unseen data.

Q: Can we predict the amount of first serves won in a match given the amount of first serves made?

A: Using linear regression, we can model the relationship between the two variables and produce an accurate answer.