DID YOU CATCH THE GAME ?

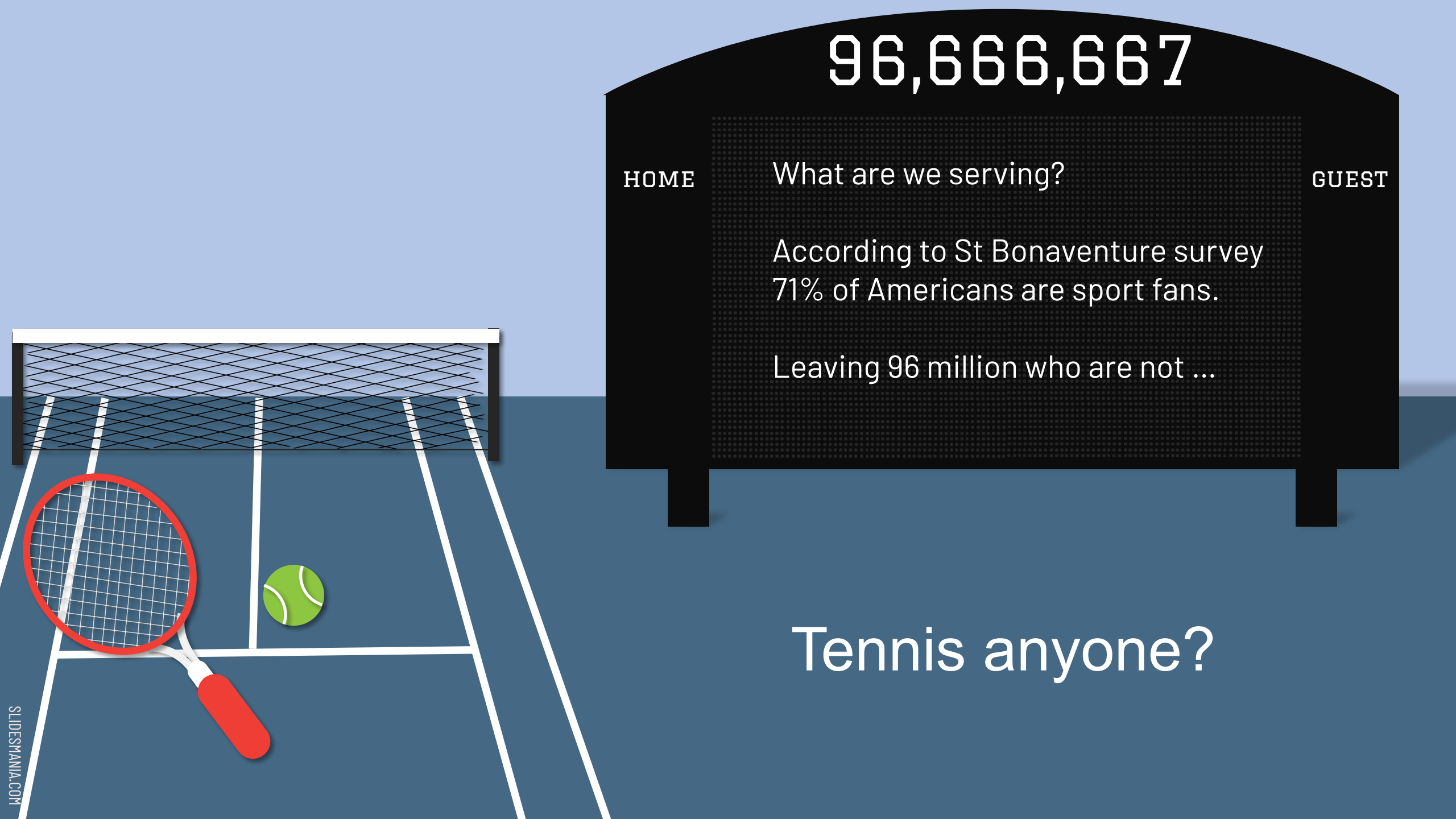
WHAT GAME?

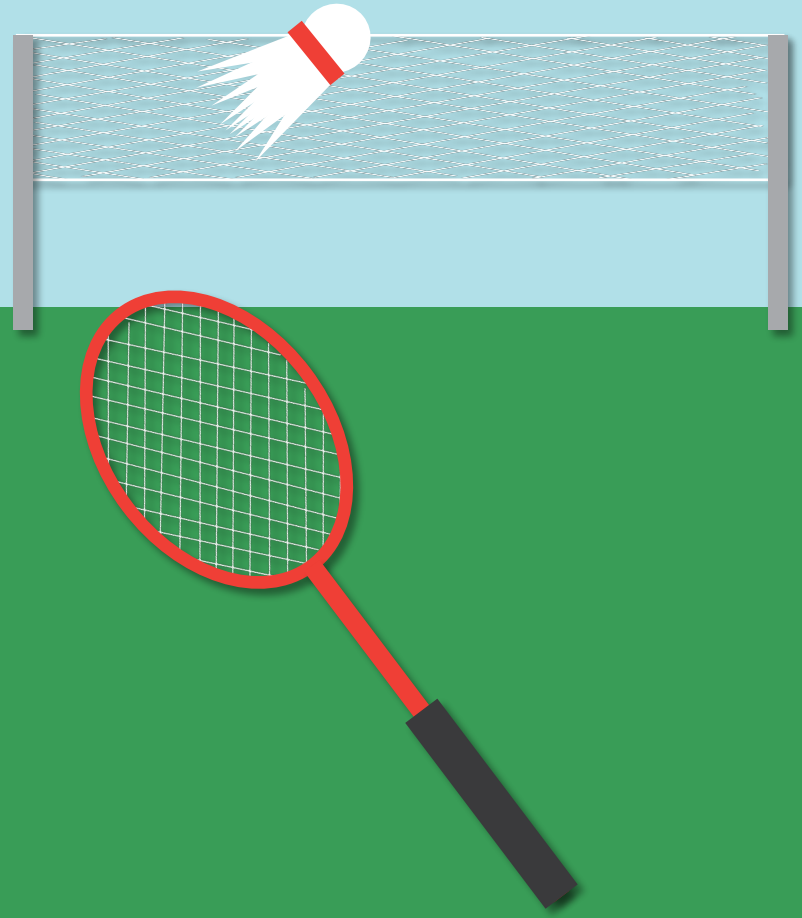
HOME

GUEST

Meet our Reddit Scraper!

Find out what is everyone talking about?


96,666,667
HOME
What are we serving?
GUEST
According to St Bonaventure survey
71% of Americans are sport fans.
Leaving 96 million who are not ...
Tennis anyone?
SLIDESMANIA.COM



3 OUT OF 10

HOME

You probably know someone who can't name a player in the game, the score of the game, or what game is being played.

Like this one?

GUEST

This is Badminton

REDDIT SCRAPER

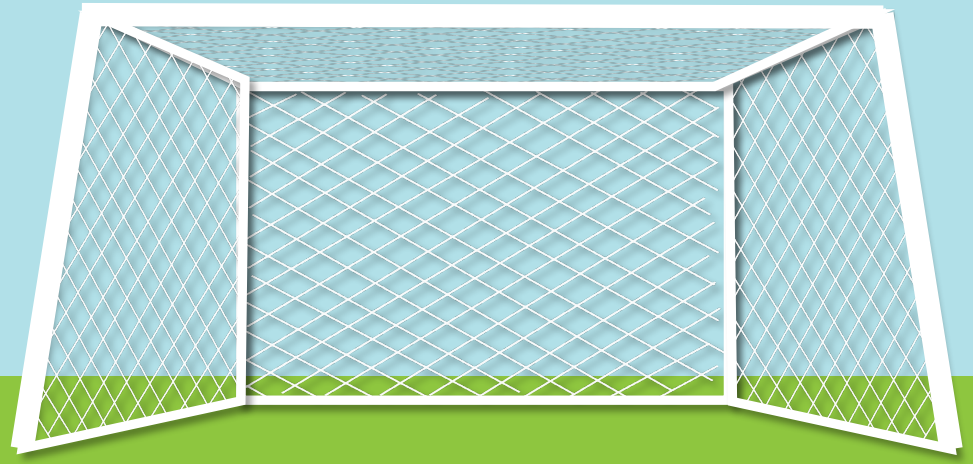
HOME

We have the tool for you!

GUEST

We have a product that can help
you study and stay relevant if
you want to up your sport lingo.

Reddit Scraper



Visit our website and follow the instructions. You can gather what people are talking about in what sports you are curious about.

Once you are comfortable, practice talking with our model. Send it a message and it will tell you if you fit in with the right crowd

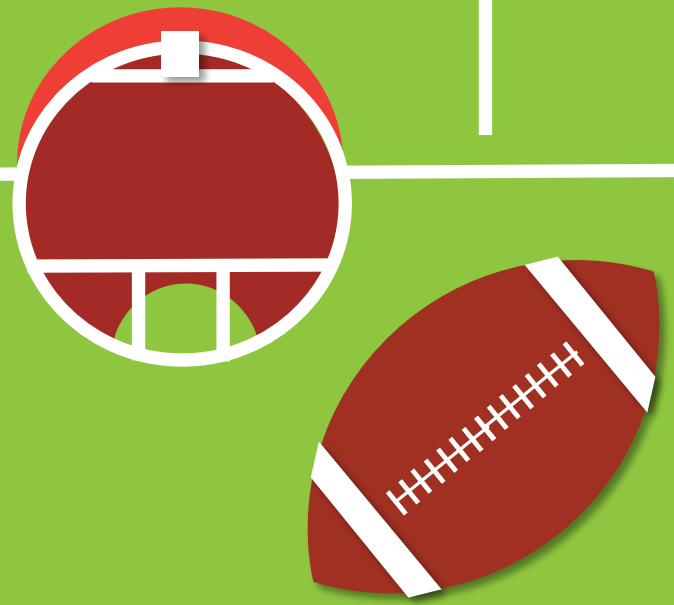
FOOTBALL

HOME

For this modeling project 1506 post titles were collected from the NFL subreddit. Another 916 were collected from the NHL.

Totalling 2422 post titles.

GUEST



There were a few hiccups at first not picking best data points and or not including some of the useful data, but eventually got sorted out.

For the data collection, focused on the full name, to ensure we had unique titles of the post.

HOCKEY

HOME

The project focused on the title of the post. This was chosen over the body as the observation was made that more unique words appeared in the title than the body.

GUEST

Before passing the post titles to the models, they were transformed via the CountVectorizer or Term Frequency Inverse Document Frequency:

- *Logistic Regression Classifier*
- *kNN Classifier*
- *Multinomial Naive Bayes Classifier*
- *Random Forest Classifer*

What's the score?

HOME

GUEST

Accuracy and F1 scores

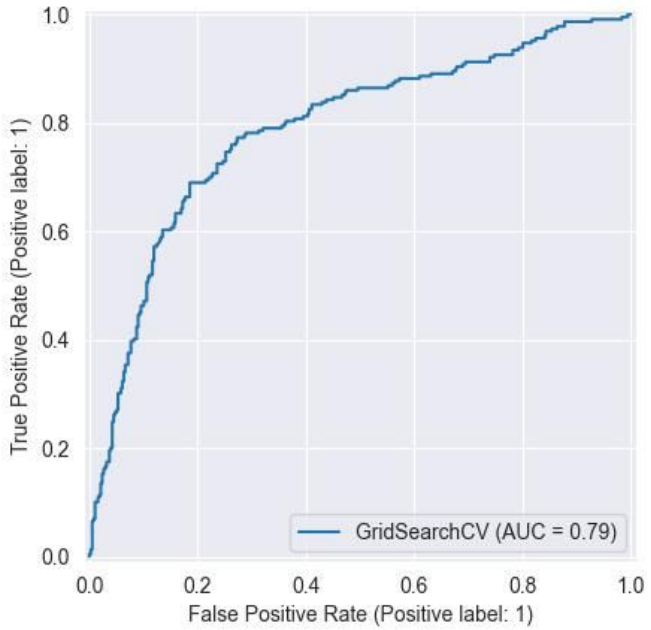
Null Model: 62.18% NFL
37.82% NHL

We want to ensure it is identifying both classes.

LOGISTIC REGRESSION

HOME

GUEST

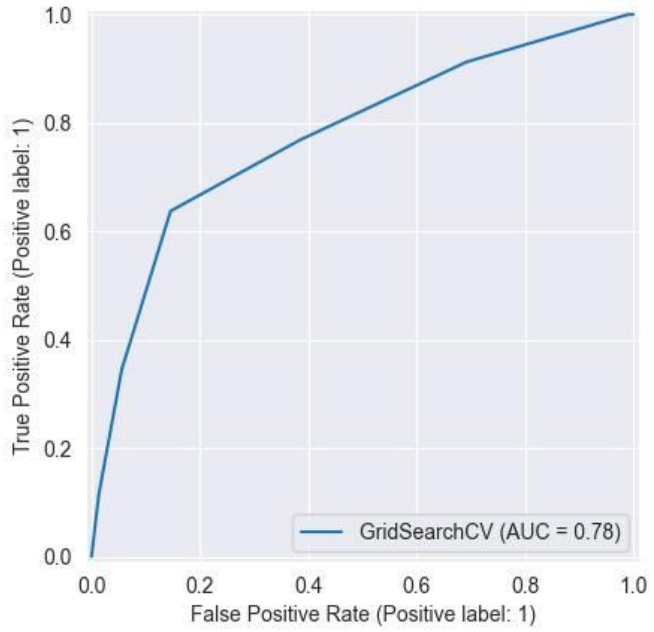


- *The CVEC model was overfit roughly 75% accurate on the testing data and 87% on the training data bearing an F1 score was 65% and 72% ROC AUC score.*
- *The TVEC model was also roughly 75% but not overfit. It did have a lower F1 score at 58% and ROC AUC at 69%*

KNN

HOME

GUEST

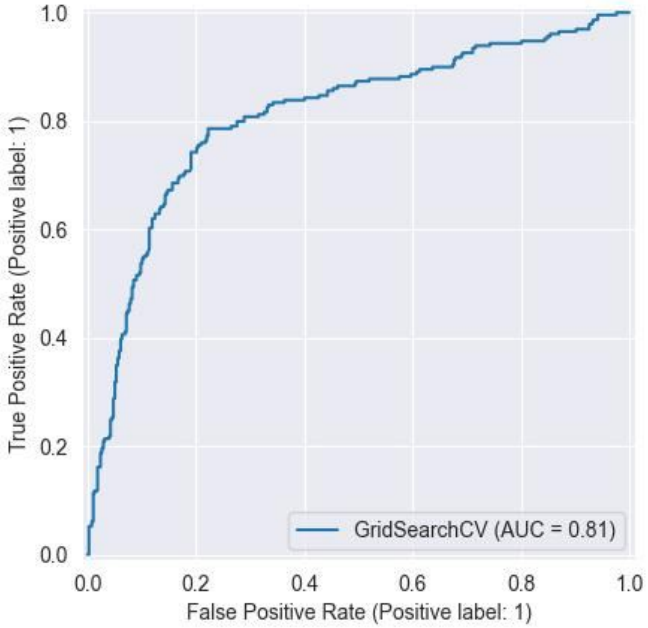


- *The CVEC model was even more overfit roughly 65% accurate on the testing data and 85% on the training data bearing an F1 score was 59% and 65% ROC AUC score.*
- *The TVEC model did better with accuracy at roughly 77% but not overfit. It did have a lower F1 score at 68% and ROC AUC at 75%*

Naive Bayes

HOME

GUEST

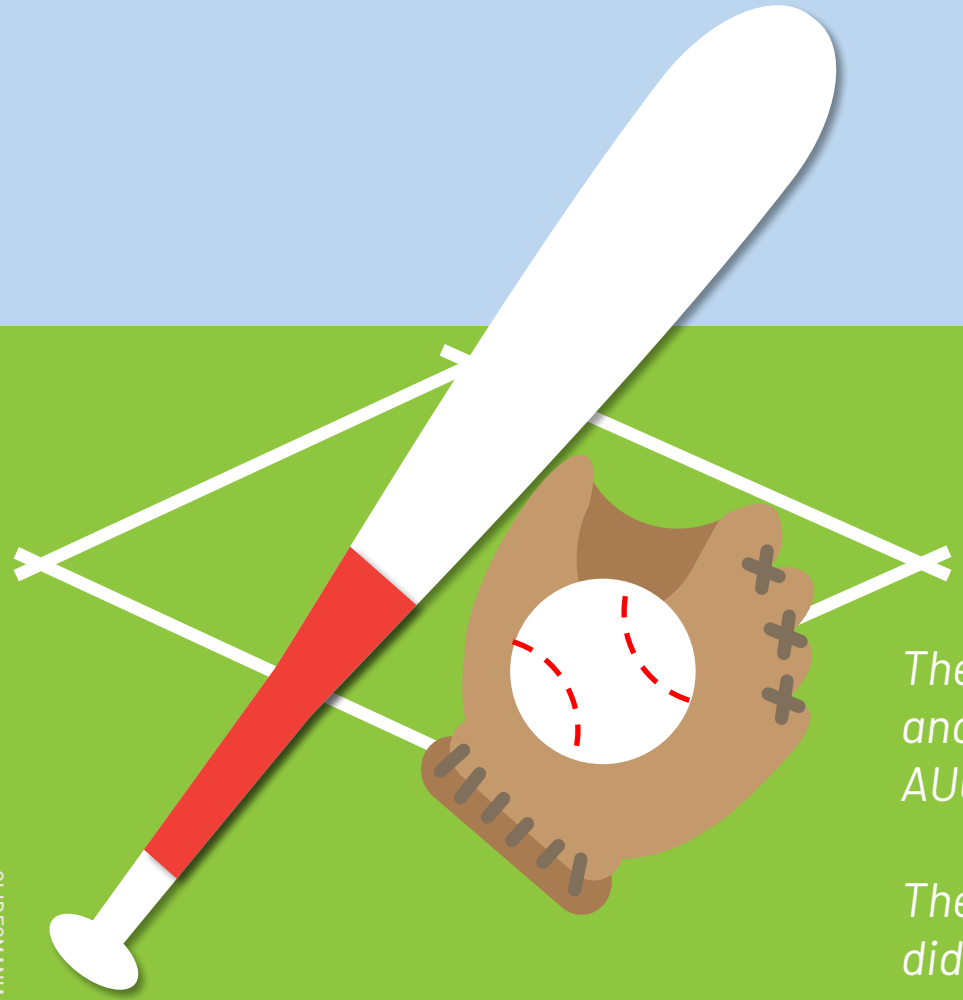


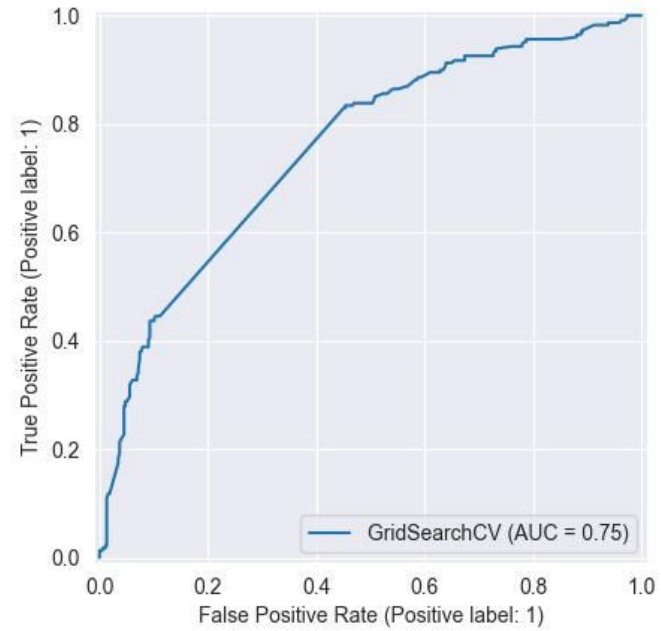
- *The CVEC model was lightly overfit roughly 78% accurate on the testing data and 86% on the training data bearing an F1 score was 69% and 75% ROC AUC score.*
- *The TVEC model did better with accuracy at roughly 77% but not overfit. It did have a lower F1 score at 66% and ROC AUC at 73%*

RANDOM FORESTS

HOME

GUEST





The CVEC model was not overfit roughly 70% accurate on the testing data and 72% on the training data bearing an F1 score was 41% and 62% ROC AUC score.

The TVEC model did better with accuracy at roughly 68% but not overfit. It did have a lower F1 score at 35% and ROC AUC at 59%

Multinomial Naive Bayes!!!

HOME

THANK YOU

Any Questions?

GUEST