Golf Sponsorship for our Client

Elliott Iturbe, Alex Valencia, Cameron DeArman

Data Cleaning Steps

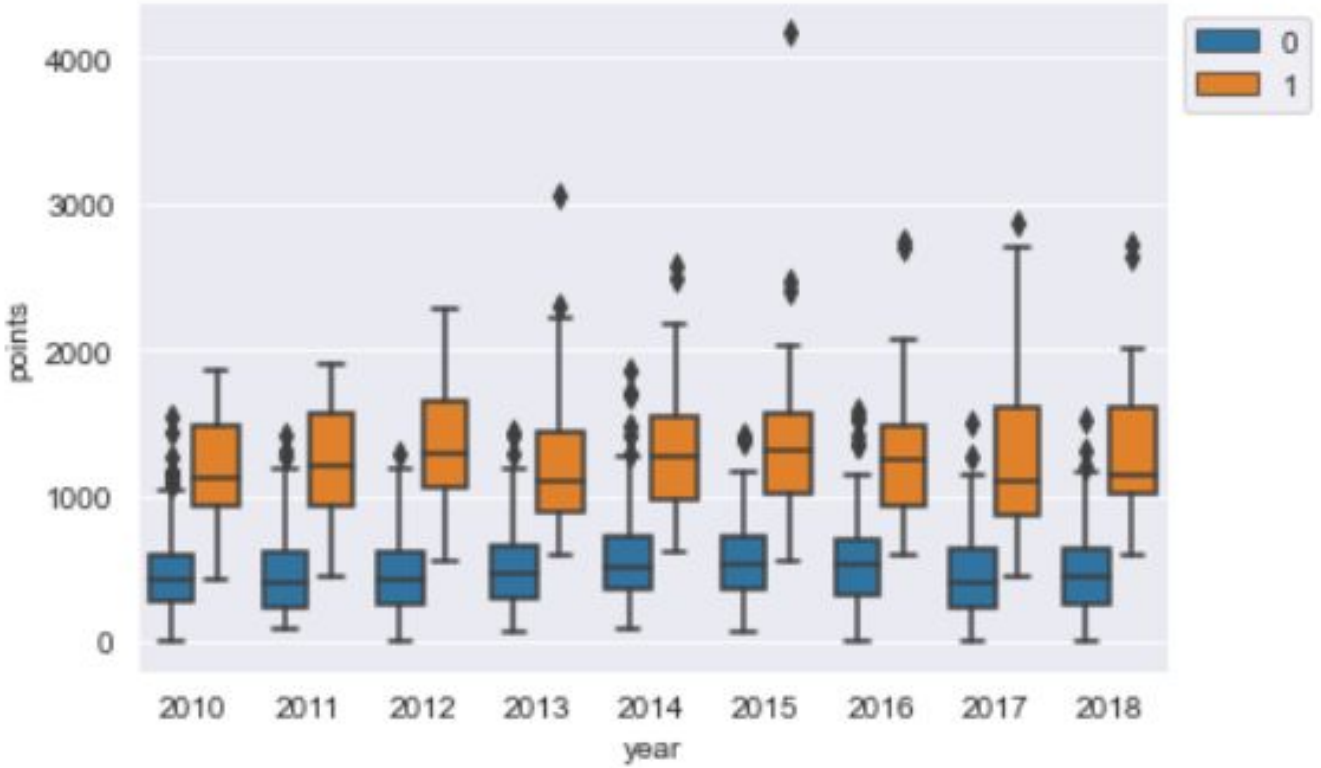
- Replaced nulls with zeros in the columns ‘top_10’ and ‘wins’
- Converted money to floats because they were objects and converted points to int because it was an object
- Replaced the question mark, commas and spaces in money column
- Changed all columns to lowercase and removed spaces

Preprocessing Steps

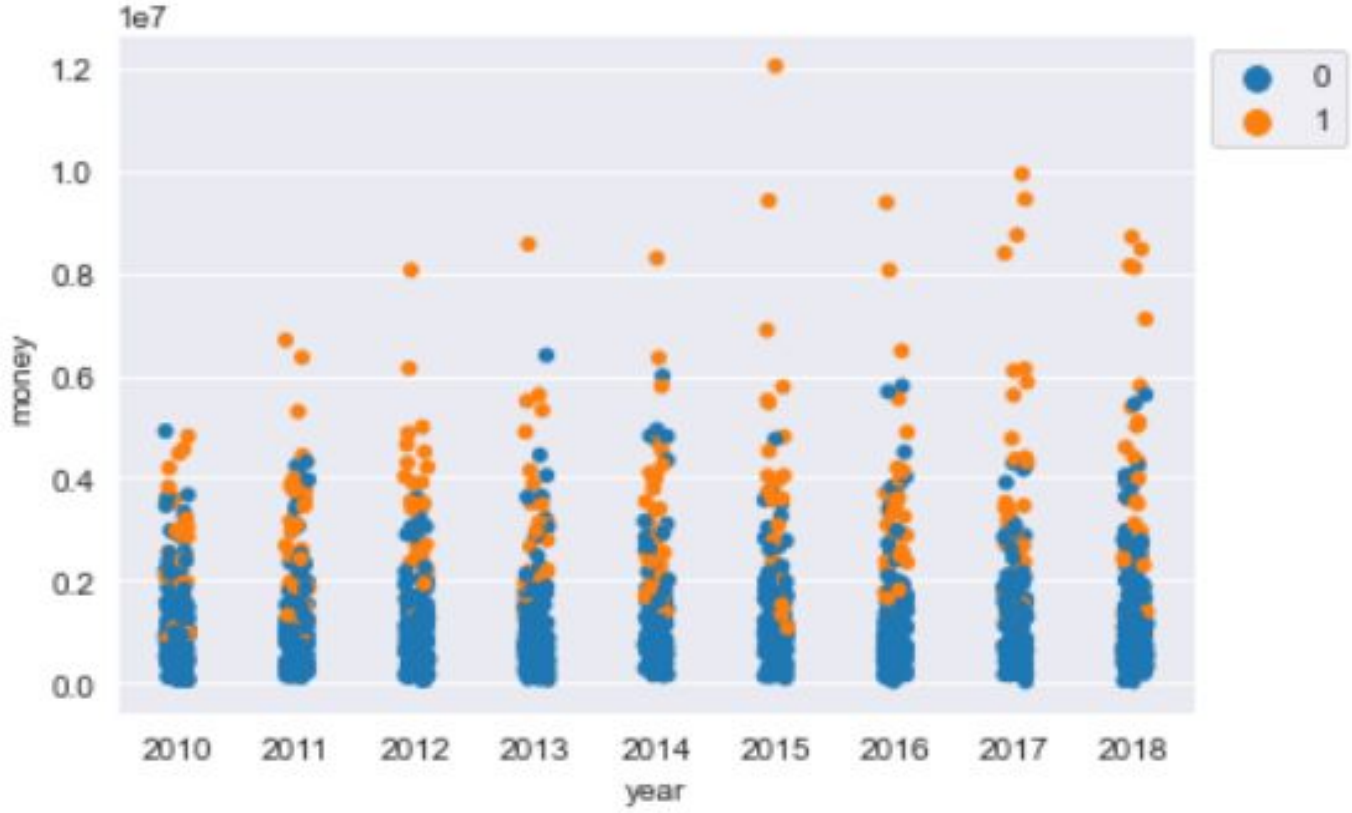
- Created a new column ‘Winners’ from ‘wins’
- Created a new column ‘Distance_fairway’ column from 'avg_distance' and 'fairway_percentage'
- Filled missing values with the median
- Scaled our data so our features were equal in weight
- Created a column ‘point_range’ that labeled the interquartile range of points with 1 and others with 0
- Created a Dataframe displaying all these changes

Points Earned by Winners and non-Winners

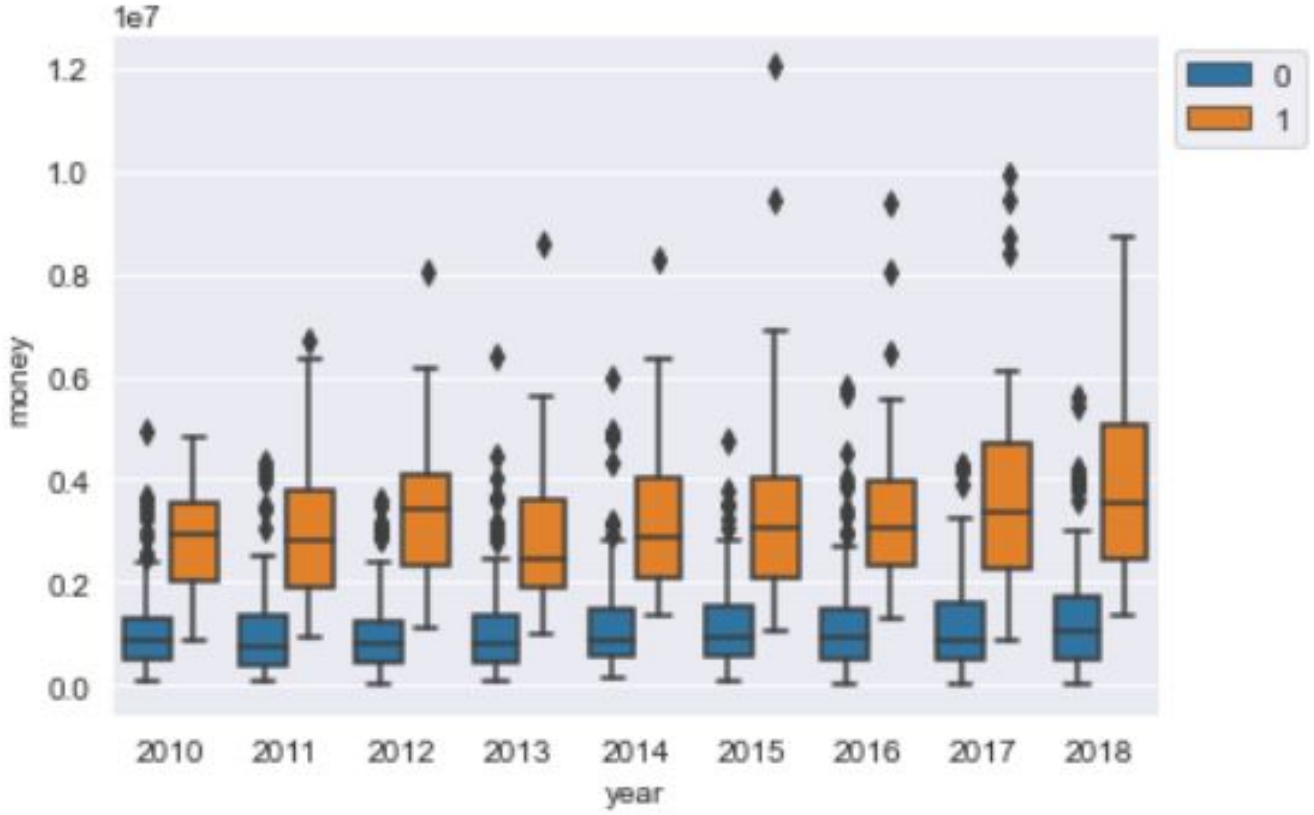
Points Earned by Winners and non-Winners



Money Earned by Winners and non-Winners



Money Earned by Winners and non-Winners



Models for Money

Baseline and Final Model for Money Entres

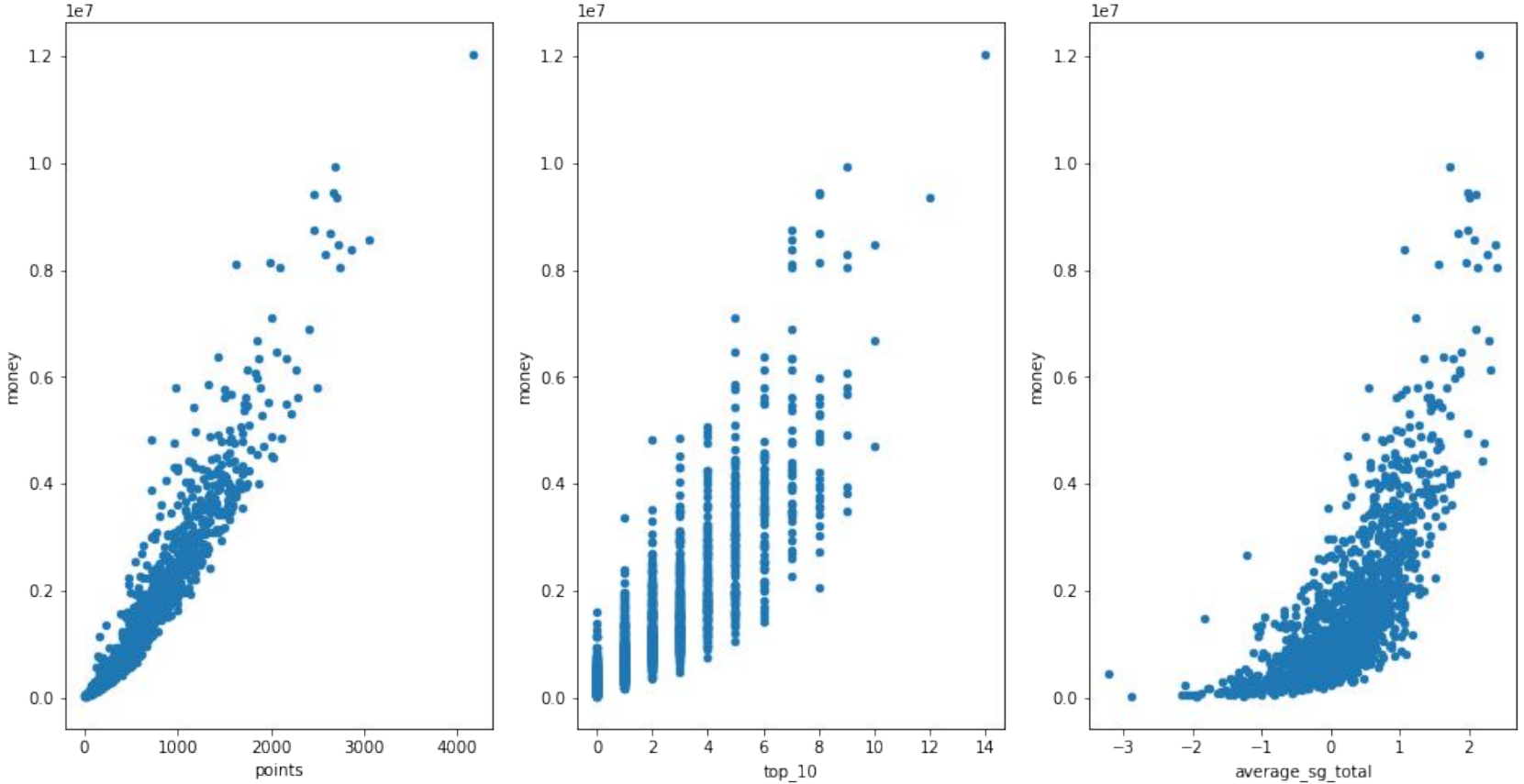
Features:

- Rounds
- Green in Regulation
- Average Putts
- Average Scrambling
- Average Score
- Points
- Top 10
- Average Strokes Gained Putts
- Average Strokes Gained Total
- Strokes Gained: Off the Tee
- Strokes Gained: Approach
- Strokes Gained: Around the Green
- Distance per Fairway
- Winners

Target

- Money

Three Most Correlated Features to Money



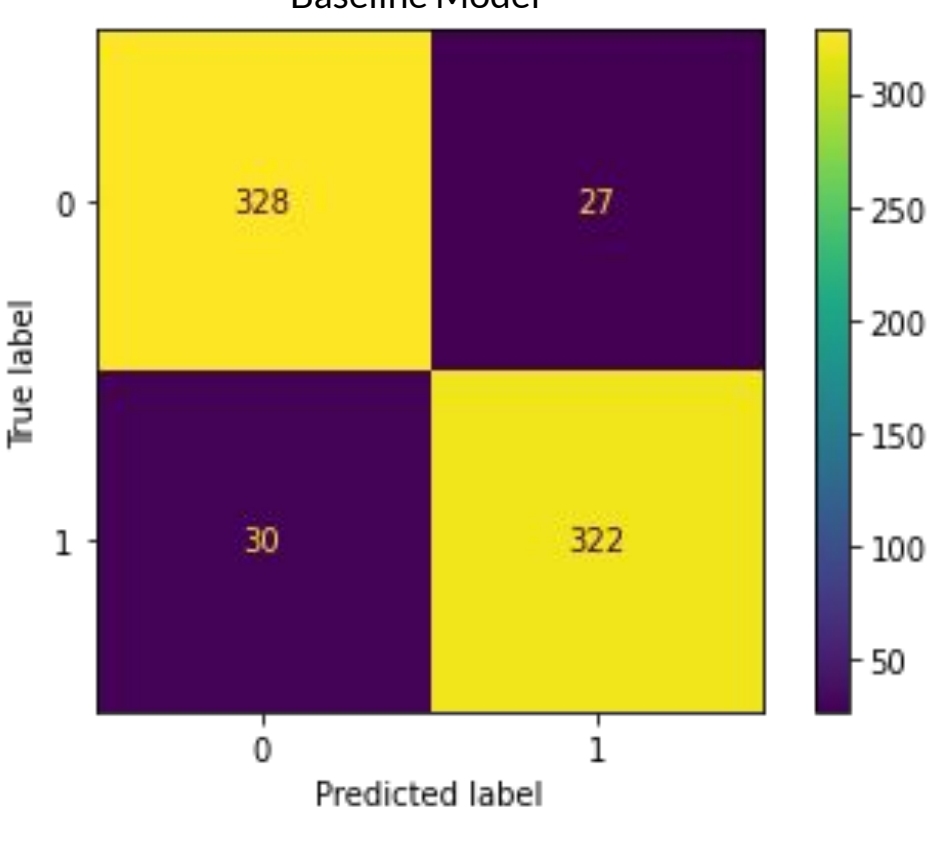
Baseline vs. Final Model Scores

- Baseline Model
  - Accuracy score:
    - Training = .919
    - Validation = .939
  - Precision score:
    - Training = .922
    - Validation = .940
  - Recall score:
    - Training = .914
    - Validation = .935

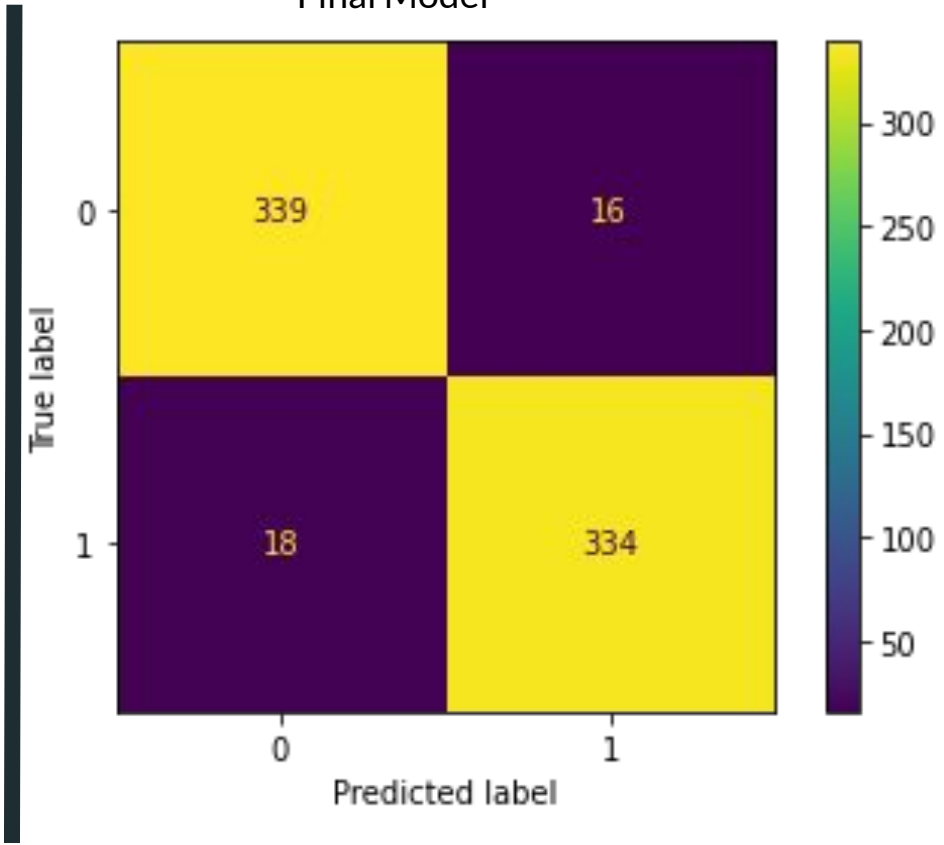
- Final Model
  - Accuracy score:
    - Training = .956
    - Validation = .955
    - Test = .921
  - Precision score:
    - Training = .965
    - Validation = .971
    - Test = .935
  - Recall score:
    - Training = .946
    - Validation = .935
    - Test = .910

Confusion Matrix

Baseline Model



Final Model



Models for Winners

Baseline and Final Model for Winners Entres

Features:

- Rounds
- Fairway Percentage
- Year of competition
- Average Driving Distance
- Green in Regulation
- Average Putts
- Average Scrambling
- Average Score
- Top 10
- Average Strokes Gained Putts
- Average Strokes Gained Total
- Strokes Gained: Off the Tee
- Strokes Gained: Approach
- Strokes Gained: Around the Green
- Money

Target:

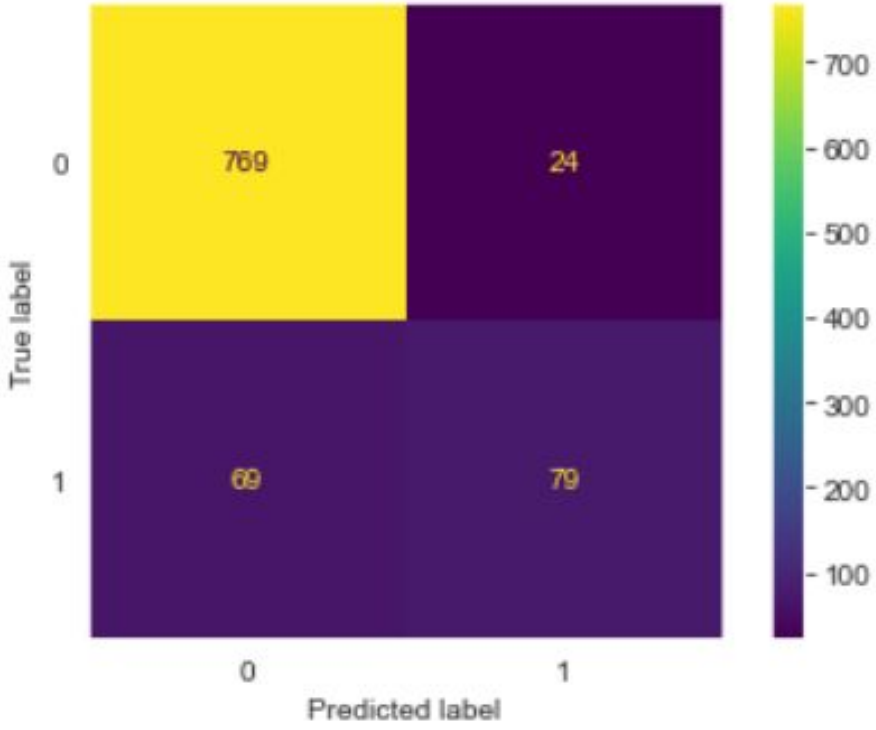
- Winners

Baseline vs. Final Model Scores

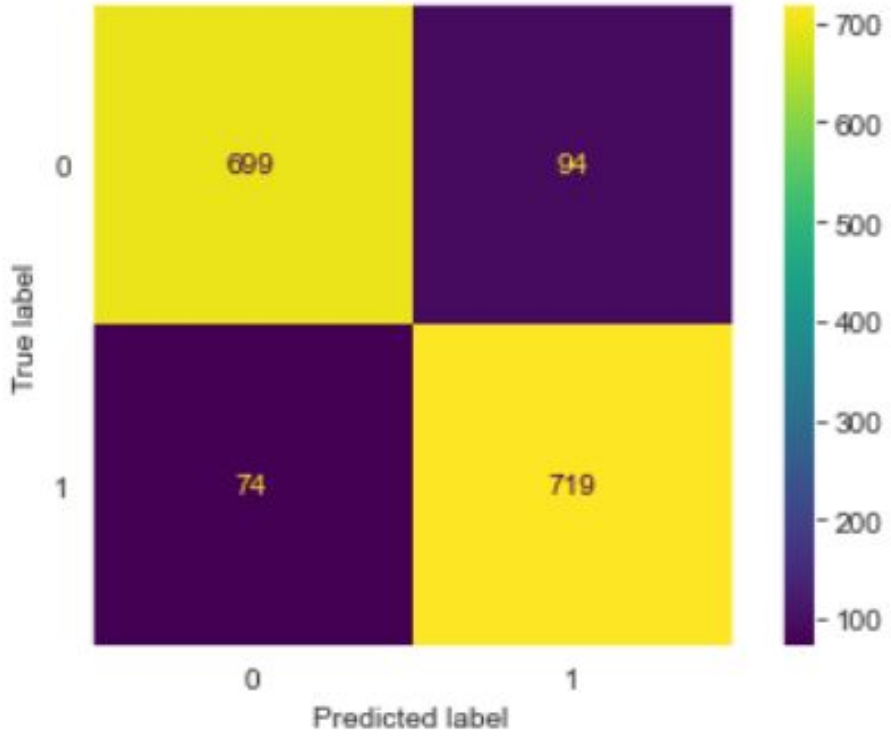
- Baseline Model
  - Accuracy score:
    - Training = .901
    - Validation = .875
- Final Model
  - Accuracy score:
    - Training = .894
    - Validation = .895
    - Test = .895

Confusion Matrix

Baseline Model



Final Model



Models on Points

Baseline and Final Model for Points Entres

Features:

- Rounds
- Green in Regulation
- Average Putts
- Average Scrambling
- Average Score
- Money
- Top 10
- Average Strokes Gained Putts
- Average Strokes Gained Total
- Strokes Gained: Off the Tee
- Strokes Gained: Approach
- Strokes Gained: Around the Green
- Distance per Fairway
- Winners

Target:

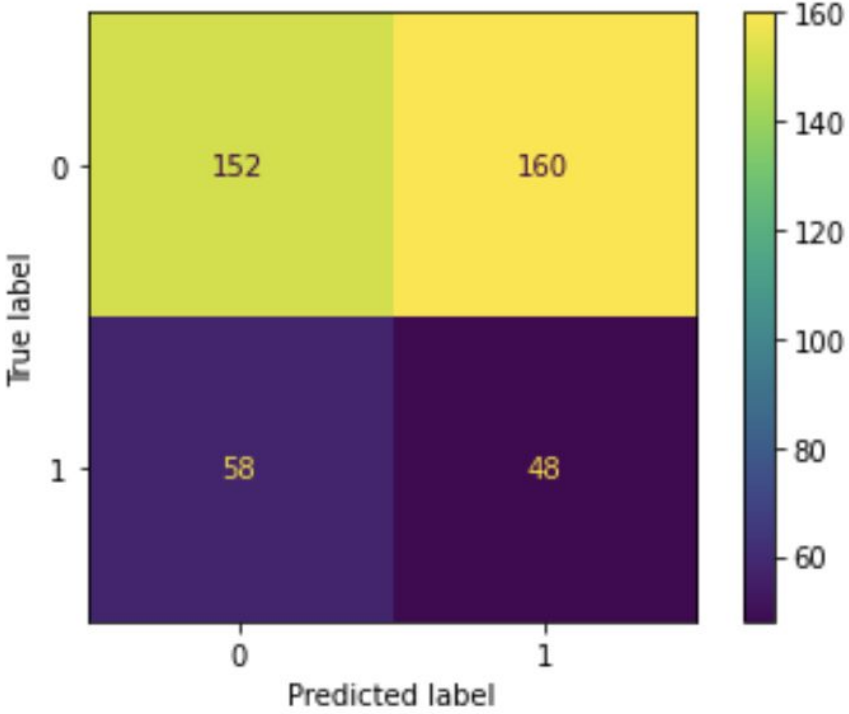
- Points

Baseline vs. Final Model Scores

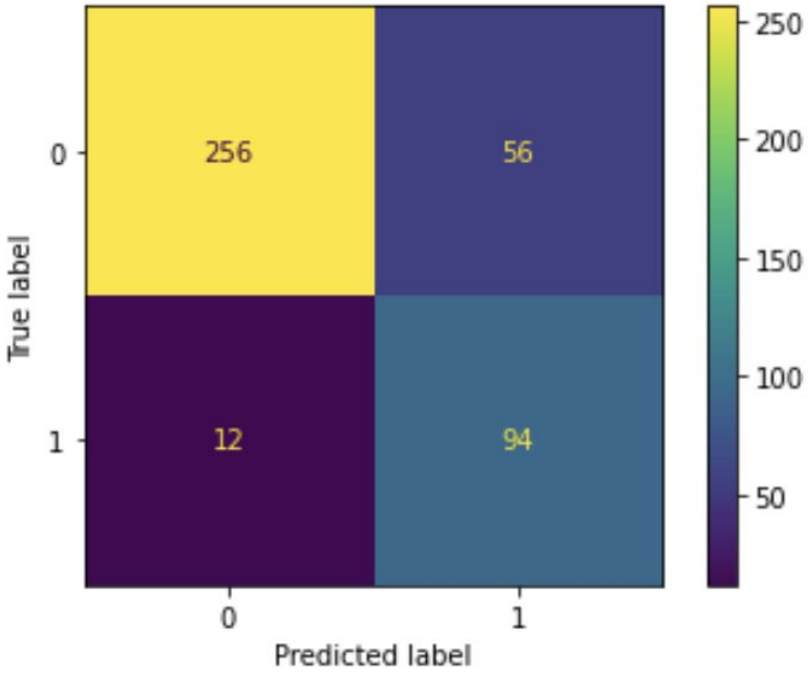
- Baseline Model
  - Accuracy score:
    - Training = .494
    - Validation = .466
- Final Model
  - Accuracy score:
    - Training = .897
    - Validation = .830
    - Test = .780

Confusion Matrix

Baseline Model



Final Model



Making the Decision

- While the models are great, there is more to the decision.
- Do the players on the list meet certain criteria outside of the models?
- Negative examples: Do they have a criminal history, any scandals or are they constantly in the media eye for bad behavior?
- Positive examples: Active in and around community, have a positive social media platform and has an overall good character in the perception of the media and the fans?
- No matter what the model tells us, it is only giving us names based off the statistics against: money, points, and winners. So there's still a decision to be made based off what's good for the business.

Thank You

Elliott Iturbe
Email: eaiturbe@bsc.edu
GitHub: @eaiturbe

Cameron DeArman
Email: cmdearma@bsc.com
GitHub: @camdeman

Alex Valencia
Email: asvalencia1688@gmail.com
GitHub: @asval211

Questions?

Index

Understanding the Columns

player_name = Name of the player.

rounds = Number of rounds completed.

fairway_percentage = This statistic refers to the number of times during the round your drive landed in the fairway (the fairway, not the light rough). It is similar to the greens in regulation, except that the maximum number per round is less than 18.

year = Year of competition.

avg_distance = The average driving distance is typically measured on two holes at each tournament and can result in nearly 40,000 shots being measured over the course of a season on some tours.

gir = Green in regulation (GIR) A green is considered hit "in regulation" if any part of the ball is touching the putting surface while the number of strokes taken is at least two fewer than par (i.e., by the first stroke on a par 3, the second stroke on a par 4, or the third stroke on a par 5).

average_putts = Average Putts is the average of putts by a player per round in the given year.

average_scrambling = Average Scrambling in golf is defined as: The percent of time a player misses the green in regulation, but still makes par or better on average based off their rounds.

average_score = Average Score is the average of all rounds played and scored by a player in a given year.

Understanding the Columns

points = FedExCup Points are awarded by finish position as defined in the point distribution tables. Limited field events during the FedExCup competition will not redistribute the points for places that do not play.

wins = The amount of times a player as won in that given year.

top_10 = Top 10 is the amount of times a player placed in the top 10 in an event of a given year.

average_sg_putts = Strokes Gained Putting reflects your performance on all putts. It compares the actual number of putts taken to the expected number of strokes to hole out based on the initial distance to the pin.

average_sg_total = Strokes gained: total simply compares a player's score to the field average. For example, a player will gain three strokes on the field if he shoots 69 on a day when the field averages 72.

sg:ott Strokes Gained: Off the Tee does just what it says, tracking a player’s performance off the Tee. This stat only applies to Par 4s and 5s, which is something to keep in mind since a solid driver of the ball can still struggle with Par 3s even if their performance is off the charts in this category.

sg:apr = Strokes Gained Approach reflects your performance on shots taken from more than 50 yards from the green, including layup shots. It takes into account the lie you were hitting from, as well as distance and accuracy.

sg:arg = Strokes gained: around-the-green measures player performance on any shot within 30 yards of the edge of the green. This statistic does not include any shots taken on the putting green.

Diagram of a Golf Hole