

TennisViz Data Scientist Vacancy Task

The aim of this task is to develop a machine-learning model end-to-end. The attached dataset contains ball and player tracking data for over 180,000 serves.

Please note: this dataset is proprietary and confidential. It can only be used for the purposes of this task and you cannot share it with anyone. You must delete the file entirely as soon as the recruitment process ends.

Each row of the dataset represents a serve and the tracking attributes of that serve (columns explained below). Each row also shows whether the serve resulted in an ace.

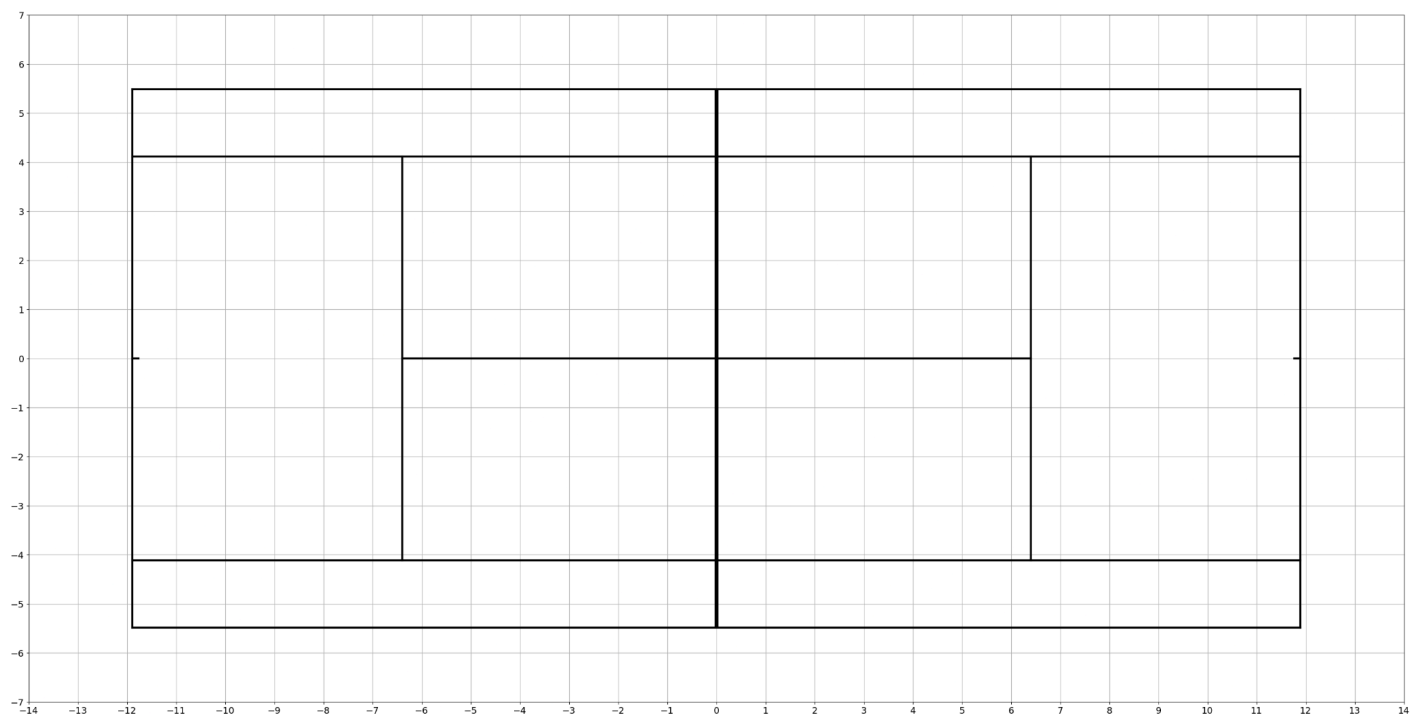
There is no additional metadata in the dataset and it has been randomly shuffled. It is therefore not possible to identify matches, tournaments, or specific player information.

Data dictionary

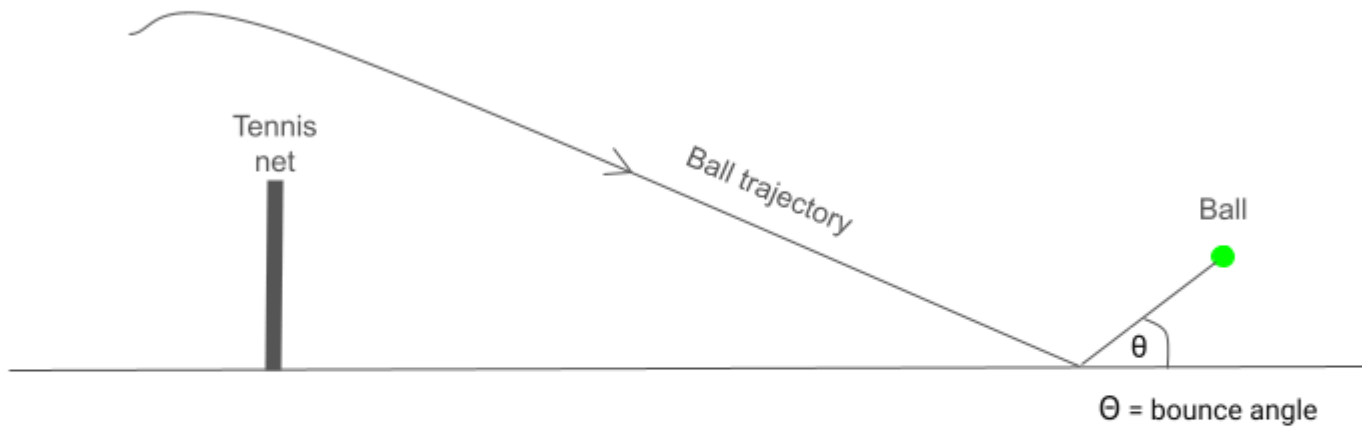
column	description
surface	one of 'hard', 'clay' or 'grass'
serve_side	one of 'deuce', 'ad'
serve_number	1 if first serve, 2 if second serve
ball_hit_y	y co-ordinate of the ball at its hit point
ball_hit_x	x co-ordinate of the ball at its hit point
ball_hit_z	z co-ordinate of the ball at its hit point (see notes)
ball_hit_v	speed (mph) of the ball at its hit point
ball_net_v	speed (mph) of the ball at the moment it crosses the net
ball_net_z	z co-ordinate of the ball at the moment it crosses the net
ball_net_y	y co-ordinate of the ball at the moment it crosses the net
ball_bounce_x	x co-ordinate of the ball bounce position
ball_bounce_y	y co-ordinate of the ball bounce position
ball_bounce_v	speed (mph) of the ball immediately after its bounce
ball_bounce_angle	angle of ball (degrees) to the ground immediately after its bounce
hitter_x	x co-ordinate of hitter at the moment the ball is hit (see notes)

hitter_y	y co-ordinate of hitter at the moment the ball is hit (see notes)
receiver_x	x co-ordinate of receiver at the moment the ball is hit (see notes)
receiver_y	y co-ordinate of receiver at the moment the ball is hit (see notes)
hitter_hand	dominant hand of hitter, one of 'right', 'left'
receiver_hand	dominant hand of receiver, one of 'right', 'left'
is_ace	1 if the serve results in an ace, 0 otherwise

Court plot with x (horizontal) / y (vertical) co-ordinates



Bounce angle diagram



Notes:

- Only serves landing 'in' are included in the dataset.
- All co-ordinates (x, y, z) are measured in metres.
- The origin (0, 0, 0) is at the base of the centre of the net.
- z co-ordinates represent height where the ground is 0; these values will always be positive.
- "bounce_angle" is always positive and can be treated as an absolute value.
- Player tracking features ("hitter_x/y", "receiver_x/y") refer to the co-ordinates of the player's centre of mass.

Task

Your task is to create a model to predict aces, using the features given to you in this dataset. When developing your model you should also think about how the model could be used in the wider context of tennis analytics. For example, it may be possible to use a model like this to evaluate the skill of certain aspects of a player's game. To do this, you may choose to intentionally omit certain features that you intend to measure.

In the subsequent interview, we will discuss the model you created, your considerations when building it, and how you may wish to use a model like this in production.

As part of the task you should try to show the following:

- Data exploration
- Visualisations
- Feature engineering
- Model selection
- Model evaluation
- Model interpretation
- Documentation

You may present your work in any number of ways, including but not limited to:

- Scripts in a private GitHub repository
- Jupyter notebook
- PDF document
- Powerpoint presentation

We strongly value clean code that is reproducible. The code you submit should be easy for us to follow, and easy for us to run end-to-end to gain the same results as you. We'll also run your model on a test set to check its accuracy. So a reproducible script to generate the model is especially important!

Finally, please spend no longer than a few hours on the task. If you're not able to complete the task within that time, that is okay! In this case, please outline in your documentation how you would intend to proceed, given more time.