Data Mining Report

Majed Naser

May 2023

1 Business Understanding

1.1 Business Objectives

Before going into details about our project vision, we will explain the identity of our client, his needs and requirements, and how we can help him reach his destination.

1.1.1 About the Client

Bill Benter is a well-known figure in the world of gambling, particularly in the horse-betting industry. With his sharp analytical mind and mathematical skills, Benter has been able to create a lucrative career out of gambling. Benter's success in the industry is evidenced by his impressive net worth, which is estimated to be around 1 billion dollars.

However, Benter's ambition to increase his profit has proven to be a challenge due to the narrow market of horse-betting in other parts of the world. Therefore, he deduced that crossing-over to other disciplines of the gambling market would be necessary, sports-betting in particular. Benter believes in the power of AI, he is determined to enter the global market and has therefore decided to ask for our help to achieve this goal.
As data mining experts, we will be working closely with Benter to analyze the global gambling market and identify opportunities for him to penetrate the market beyond the confines of the United States.

1.1.2 Business Objective

The objective of our data mining project is to provide Bill Benter, a professional gambler with a focus on horse-betting, with betting advice that can help him gain a strong entry into the Tennis Gambling Market. The ultimate goal is to increase his revenues by expanding his business to the tennis-betting market and leveraging our expertise in data mining and predictive analysis.

Our ultimate goal is to increase his revenues by providing him with accurate and reliable betting advice on tennis matches.

1.1.3 Secondary Objectives

During our initial interaction with the client, we discussed several secondary business objectives that could be beneficial to his quest. One of these objectives is to provide advice on which bookmakers in the market are shady or try to keep the gambling terms blurry. This information can help the client make more informed decisions about which bookmakers to work with and avoid any potential issues that may arise from working with untrustworthy bookmakers. By providing this additional advice, we aim to help the client build a more successful and sustainable business in the long term.

Another secondary business objective that we discussed is to provide advice regarding extra betting terms. These terms may include the number of hits or misses for each player, which can be valuable information for professional gamblers who specialize in tennis betting. By providing this advice, we aim to help the client make more informed decisions about which bets to place and maximize his chances of winning.

While these secondary business objectives may not be essential to the overall success of our project, we believe that they can be highly beneficial to the client. By providing additional advice on bookmakers and extra betting terms, we aim to help him expand his business to the tennis-betting market with greater confidence and achieve his ultimate goal of increasing his revenues.

1.1.4 Business Success Criteria

The primary success criteria for our business objective will be providing useful betting advice that can guarantee a strong entry for our client into the global market. Our ultimate goal is to help the client increase his revenues by providing him with reliable and accurate advice on tennis matches. Therefore, the success of this objective will be determined by the amount of revenue earned once the client goes through a few initial bets.

For our secondary business objective of providing advice on which bookmakers in the market are shady or try to keep the gambling terms blurry, the success criteria will be measured by the effectiveness of our advice. In other words, The success of this objective will be determined by the client's success in building successful relationships with reliable and trustworthy partners with which a long-term business can be maintained. This can be initially highlighted by the number of issues or disputes that our client experiences (or doesn't experience) with bookmakers after attending our advice. If our advice helps him avoid issues with untrustworthy bookmakers, this objective will be considered successful.

The success criteria for our secondary business objective of providing advice on extra betting terms, such as the number of hits or misses for each player, will be measured by the added value that this advice provides to our client. The success of this objective will be determined by the number of bets that the client places based on our recommendations for these extra betting terms and the profitability of these bets. If our advice helps him increase his revenues and maximize his chances of winning, this objective will be considered successful.

Overall, the success criteria for our primary and secondary business objectives will be measured by the value that we provide to the client in terms of increased revenues, improved profitability, and a more sustainable business model. We will constantly evaluate and measure our performance against these success criteria to ensure that we are delivering the best possible results for our client.

1.2 Situation Assessment

In this section, we will highlight all the requirements of our project, the constraints we need to comply with, and the resources we will be working with.

1.2.1 Human Resources

Unfortunately, the workforce currently behind this project is limited. The entire workforce is made up of a few individuals from both our side and the client's and are all listed below.

Bill Benter: As the client and project profiteer, Benter will be the primary point of contact for our team. He will communicate directly with our project manager, Majed Naser, to provide input on the project vision and business objectives.

Majed Naser: As the project manager and plan initializer, Majed will be responsible for overseeing the entire project workflow and ensuring that all team members are aligned with the project vision and objectives. He will work closely with the client to ensure that our predictions align with his business objectives and will be responsible for presenting the results to Bill at regular intervals.

Mahmoud Mousatat: As our modeling expert, Mahmoud will be responsible for building our machine learning models. This will include tasks such as feature engineering, model training, testing, and tuning. He will work closely with Timur Aizatvafin to ensure that the data is suitable for modeling and will work with Majed to ensure that our models align with Benter's business objectives.

Timur Aizatvafin: As our data expert, Timur will be responsible for all data operations, including loading, cleaning, analysis, and preprocessing. He will also work closely with Mahmoud in the feature engineering step to ensure that our models are built using the most relevant and accurate features. Timur will work with Majed to ensure that our data operations align with Benter's business objectives and that our predictions are based on the most accurate and relevant data.

1.2.2 Machine & Software Resources

Based on the information provided, the main machine and software resources needed for our project will be:

Computing resources: Since the client does not have any computing resources, we will rely on our own machines to perform the data mining tasks. This means that our team members will need to have access to high-performance computing resources with enough processing power to handle the volume of data we are working with.

Open source dataset: As the client does not have any resources regarding tennis matches, we can only rely on the dataset provided to us. This also means that we need to ensure that this dataset is of good-quality and contains accurate information.

Data mining tools: Our tool stack will consist of common data mining tools such as Jupyter Notebooks, Python libraries for data manipulation and analysis, machine learning frameworks, and visualization tools. These tools will help us clean, analyze, and preprocess the data, as well as to build and evaluate machine learning models later.

1.2.3 Project Assumptions, Requirements & constraints

The project requirements, conditions, and constraints are as follows:

Non-negative profit: The primary requirement of our project is that it should not cause losses to the customer. This means that our betting advice should be accurate enough to ensure that the customer's profits are non-negative. While the customer can take advice from other sources, our advice should be properly integrated and weighed against other sources to ensure the best possible outcome.

Limited budget: The project budget is limited, which means that any data we use for our data mining project must be obtained with minimal financial costs. Any data that requires payment or additional costs must be

approved by the customer before being used.

Data consent: We must obtain consent before using any data found by our team for our data mining goal. This means that we must ensure that all data used in our project is legal, ethical, and obtained through appropriate means.

Time constraint: We have a strict time constraint to complete our project before the start of the French Open Tournament on May 21st. This means that we must ensure that all project tasks are completed within the given time frame to deliver accurate and reliable predictions to our customer.

Overall, the success of our project will depend on how well we can meet these requirements, conditions, and constraints. We must ensure that our betting advice is accurate and reliable, while also staying within budget, obtaining necessary consent for data use, and delivering our project within the given time frame.

The only assumptions our project is built upon is that tennis match results are not random or independent variables but can rather be predicted based on other variables calculated beforehand.

1.2.4 Risks & Contingencies

This section is dedicated to discuss the risks implicit in our work and our strategy in case such issues take place.

A risk that we cannot ignore is the possibility of the model being overfit to the training data, meaning it performs well on the training data but poorly on new, unseen data. To minimize this risk, we will use techniques such as cross-validation, regularization, and ensembling to ensure that the model generalizes well to new data.

Another risk that we should be aware of is failing to satisfy the deadlines for the project or one of its milestones. In such a case, we will have a dedicated team member assigned the task of online support for our customer. The support person will help the customer through data analysis tools until the model is in the deployment stage.

Finally, there is a risk of the client misinterpreting the results and relying solely on the model's advice without considering other factors. To avoid this risk, we will provide clear and concise explanations of the model's predictions and limitations. We will also encourage the client to consider other factors, such as their own knowledge, experience, and intuition, in making their betting decisions.

While these risks are not to be taken lightly, we are confident that we have the expertise and experience to manage them effectively. By being transparent about these risks and our mitigation strategies, we hope to establish a strong and trustworthy long-term partnership with our client.

1.2.5 Terminology

In this section, we share some of the basic terminology of tennis sport. This section is necessary to have a thorough understanding of the accurate meaning of our data (both for the client and the contractor).

Betting Terminology .

Bookmaker: A bookmaker is a company or individual that facilitates betting on various events, such as sports, by suggesting and accepting bets from customers. They earn money through commissions charged on each bet placed.

Bet: A bet refers to the amount of money that an individual gives to a betting company in the hope of correctly predicting the outcome of a particular event. It is a financial wager made by the customer, which is at risk with the possibility of losing the bet or winning a return.

Bet coefficient: The bet coefficient, also known as the odds or payout ratio, represents the amount of money a customer would receive as a return for each unit of money wagered. For example, if the bet coefficient is 1.05, a successful bet of 100 would result in a return of 105 (before tax) to the customer.

ROI (Return on Investment): A financial measure used to assess the profitability and success of a project or investment. It indicates the percentage or ratio of profit generated relative to the initial investment. An ROI of 1.05 would typically be considered a moderate level of success, implying a return of 5% over the initial investment.

$$ROI = \frac{NetReturnofInvestments}{CostofInvestments} \times 100\%$$

Tennis Sport Terminology .

Match: In tennis, a match is a competition between two players (singles) or two teams (doubles). It consists of several sets, and the player or team wins the match by winning a specified number of sets, usually two or three. Matches are commonly structured as "best of three" or "best of five" sets, meaning the player or team that wins the majority of sets is declared the winner.

Set: A set in tennis is a unit of scoring within a match. To win a set, a player must win a certain number of games. Typically, a player needs to win six games, with a margin of at least two games. However, if the score reaches 5-5, the player must win seven games to secure the set. In the event of a 6-6 tie, a tiebreaker is played, and the player who reaches seven points first, with a margin of two points, wins the set.

Game: A game in tennis is a smaller unit of scoring within a set. Each game consists of a sequence of points, with the serving player alternating between games. The scoring system for a game follows the sequence of 0-15-30-40, and the player must win the next point after reaching 40 to win the game. When both players reach a score of 40, a "Draw-Less-Greater" system, also known as "deuce," is applied, requiring a player to win two consecutive points to win the game.

Grand Slam Tournament: A Grand Slam Tournament is a series of prestigious and significant tennis events that are considered the most prominent in both the ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) tours. The Grand Slam tournaments include the Australian Open, French Open, Wimbledon, and the US Open. They are recognized as

the highest level of competition in professional tennis.

Walkover: In tennis, a walkover occurs when a match is finished but not played entirely due to one player leaving the match for a reason other than injury. This can happen if a player withdraws, is disqualified, or fails to show up for the match. In a walkover, the opposing player or team is awarded the victory without having to complete the match.

Retired: In tennis, when a player is unable to continue playing during a match due to injury, they are said to have retired. In such cases, the match ends prematurely, and the opposing player or team is typically declared the winner.

Backhand: The backhand is a shot in tennis executed from the left side of the player (for right-handed players) or the non-dominant side. It can be performed using either one hand or both hands on the racket, depending on the player's style and preference.

Forehand: The forehand is a shot in tennis executed from the right side of the player (for right-handed players) or the dominant side. It is one of the primary strokes in tennis and is typically performed with a single hand on the racket.

ATP: ATP stands for the Association of Tennis Professionals, which is the governing body for men's professional tennis. It organizes and oversees various tournaments and rankings for male tennis players worldwide.

WTA: WTA stands for the Women's Tennis Association, which is the governing body for women's professional tennis. It is responsible for organizing and regulating tournaments and rankings for female tennis players globally.

Data mining Terminology .

Machine Learning Model: A machine learning model is a computer program or algorithm that can analyze and process data to make predictions or decisions in a specific domain. It utilizes patterns and relationships within

the data to learn and generate insights or forecasts. For instance, a machine learning model can predict the outcome of a tennis match or estimate the age of a championship winner based on historical data.

Data Mining: Data mining refers to the process of collecting and preprocessing data from various sources in order to extract valuable information and patterns. It involves techniques such as data cleaning, integration, transformation, and selection to prepare the data for analysis and modeling. Data mining aims to discover hidden patterns, relationships, or insights that can be used for decision-making or developing predictive models.

Accuracy: Accuracy is a measure of the performance or quality of a classification system or predictive model. It quantifies how well the system can predict or classify the desired outcomes. It is commonly expressed as a percentage or a decimal value between 0 and 1. For example, if a system correctly predicts the winner of a match in 50% of cases, the accuracy would be 0.5 or 50%. A higher accuracy value indicates a better predictive performance.

1.3 Data Mining Goals

In this section, we will be viewing our goals from a data mining perspective.

1.3.1 Data Mining Goals

(1) Our project aims to examine, analyze, and explore the data from various angles, including essential tennis statistics, leading players, betting odds for matches, and various types of available bets.

(2) We will pass the dataset through TPOT, which will automate the search for the best possible pipeline to fit our data.

(3) We also intend to develop other machine learning models that can anticipate the likelihood of one player prevailing over another based on match initial data, player statistics, and additional information produced during data processing stage. The TPOT model will be

assessed in comparison to these models.

(4) As a secondary objective, we plan to publish the generated model on the cloud, enabling users to predict match scores from a website interface (This will only be done if it's later found to be in accordance with the time frame we have).

1.3.2 Data Mining Success Criteria

Naturally, some Data Mining Success Criteria was chosen in convenience with our Data Mining Goals. This criteria is as follows:

– A Comprehensive Analysis is done, a meaningful output of plots tennis statistics, player information, betting odds, and types of bets, are provided, after thorough examination and analysis.

– A Machine Learning Model is developed, capable of presenting winning probabilities for both players.

– The model achieves an accuracy higher than 60%, which would be enough to provide a moderately high sustainable profit.

– Model Deployment (if feasible): If within the project's time frame, our chosen model should be successfully deployed to the cloud, allowing users to access and utilize the model's predictions for match scores through a user-friendly interface.

1.4 Project Plan

In this section, we present our initial plan being in action during the timeline of three months, handling the stages of our Data Mining Project by their natural order.

1.4.1 Project Plan in Iterations

The project workflow was split into several iterations. Each of them almost corresponding to a stage (or stages) of the CRISP-DM stages.

– First Iteration: This iteration involved learning the tennis terminology, learning the true meaning of the features of the ATP Tennis Dataset, constructing the business objectives that can be served using this dataset, and finally formulating the Data Mining goals in accurate terms. This iteration was mostly concerned with the Business Understanding stage of our report. This iteration also involved frequent communication with the client to help build a foundation of common views that can help build an ideal project plan.

– Second Iteration: This iteration involved exploring, analyzing, and visualizing the data, The non-relevant features were dropped, some relevant ones were used to construct more features. At the end of this stage, an adequate model-training dataset was built.

– Third Iteration: This iteration involved building, training, and fine-tuning the actual model. it also included searching for the best possible model via TPOT, and then evaluating the model with regards to its feasibility to our success criteria. It was also during this iteration that our presentation was built and recorded, and our report was finalized.

1.4.2 Initial Assessment of Tools and Techniques

– Business Understanding: CRISP-DM, Customer Communication.

– Data Understanding: Tableau.

– Data Preparation: Jupyter notebooks (python), MS Excel.

– Modeling & Evaluation: Decision Trees, TPOT, ROI

– Deployment: Dashboard.render.com, telegram bot.

2 Data Understanding

2.1 Data Collection

To facilitate our comprehensive analysis of the tennis industry, we were provided with a dataset containing valuable information on men's tennis games dating back to the year 2000. This dataset plays a crucial role in our data understanding process as it enables us to identify significant trends and patterns in player performance, tournament outcomes, and more.

Of particular significance is the matches.csv file, which provides an extensive overview of thousands of matches held over the past two decades. This rich dataset empowers us to delve into player rankings, tournament results, and other essential metrics, offering profound insights into the present state of men's tennis. (Note that the dataset exclusively pertains to men's tennis and does not encompass women's tennis data.)

By analyzing this data carefully, we hope to gain insights into which factors contribute to successful tennis players, which strategies are most effective in winning matches, and how the tennis industry has evolved over time. Overall, these files will be an essential resource as we dive deeper into our analysis of men's tennis.

Enclosed herewith are comprehensive tables that provide a detailed breakdown and explanation of each individual feature present within the dataset tables. These tables serve the purpose of enhancing understanding by providing clear and extensive insights into the characteristics, attributes, and variables encompassed by the dataset.

2.2 Data Description

Table 1: Matches

Field	Description
ATP	Tournament number
Location	Venue of tournament
Tournament	Name of tournament
Date	Date of match
Series	Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
Tier	Tier (tournament ranking)
Court	Type of court (outdoors or indoors)
Surface	Type of surface (clay, hard, carpet or grass)
Round	Round of match
Best_of	Maximum number of sets playable in match
Winner	Match winner
Loser	Match loser
WRank, LRank	ATP ranking of the match winner/loser as of the start of the tournament
WPts, LPts	ATP points of the match winner/loser as of the start of the tournament
W1-W5	Number of games won in each of the sets by match winner
L1-L5	Number of games won in each of the sets by match loser
Wsets, Lsets	Number of sets won by match winner/loser
Comment	Match status (completed, won through retirement of loser, or via walkover)
B365W,...UBW	11 bookmaker coefficients for match winner
B365L,...UBL	11 bookmaker coefficients for match loser
MaxW, MaxL	Maximum coefficient for match winner/loser
AvgW, AvgL	Average coefficient for match winner/loser

Table 2: Players

Field	Description
Player_ID	Unique identifier for the player
First_name	First name of the player
First_initial	First initial of the player's name
Last_name	Last name of the player
Full_name	Full name of the player
Player_URL	URL to the player's profile at the official website of Men's Professional Tennis
Flag_code	Three-letter abbreviation country code of the player
Residence	City of the player's current residence
Birthplace	Country and city of the player's birth
Birth, year, month, day	Date of birth of the player
Turned_PRO	Year when the player became a professional player
Weight (kg, lbs)	Weight of the player in kilograms and pounds
Height (cm, inches)	Height of the player in centimeters and inches
Handedness	Dominant hand of the player (left or right-handed)
Backhand Hit in tennis	Type of backhand stroke in tennis (one-handed or two-handed)

2.3 Data Exploration

Popular Court: This graph shows the distribution of tennis matches played on different court surfaces since 2000. It can help you identify which court types are the most popular among players and how this popularity has changed over time.

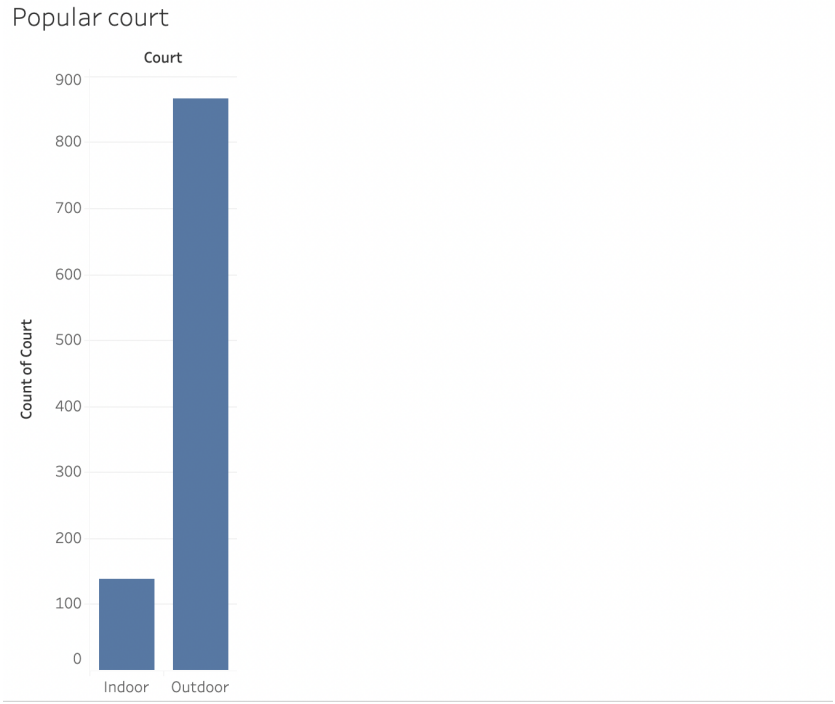


Figure 1: Popular Court

How Many Times Players Lose in a Year: This graph displays the number of losses for the top 10 tennis players in each year since 2000. It can help you identify which players have been the most successful over the years and how their performance has changed over time.

Popular Places for Tennis Matches: This graph shows the distribution of tennis matches played in different countries since 2000. It can help you identify which countries are the most popular for hosting tennis tournaments and how this popularity has changed over time.

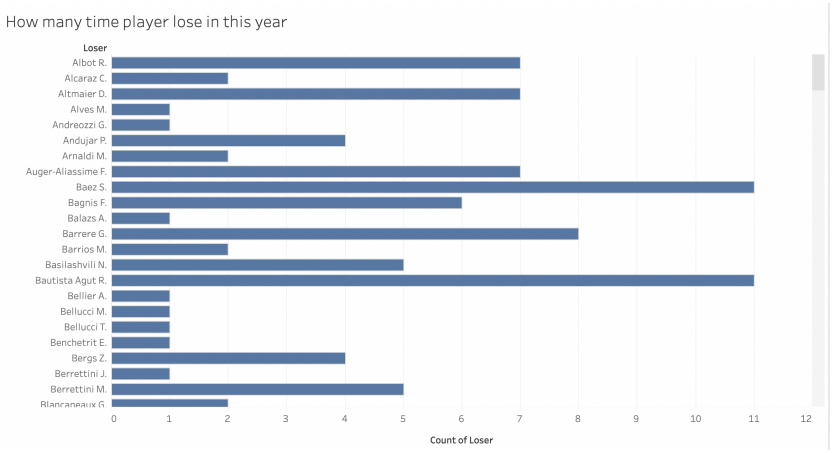


Figure 2: How Many Times Players Lose in a Year

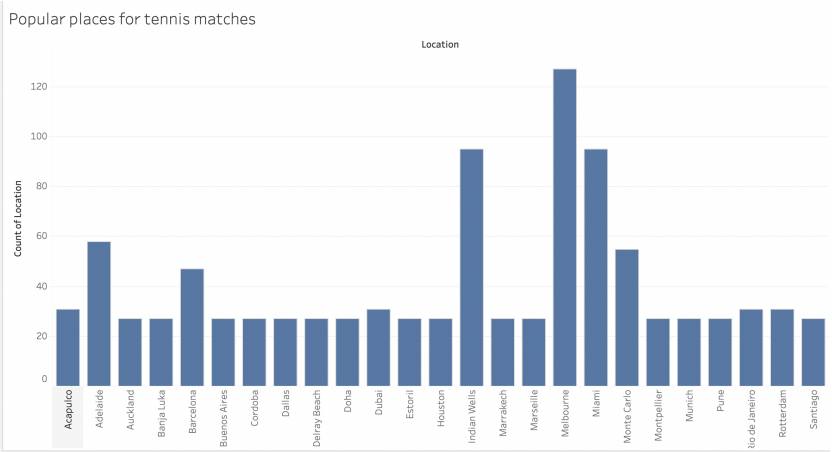


Figure 3: Popular Places for Tennis Matches

3 Data Preparation

3.1 Data Selection

3.1.1 Matches

Field	Reason for Dropping
Comment	No useful info
CBW	A lot NaN
CBL	A lot NaN
GBW	A lot NaN
GBL	A lot NaN
IWW	A lot NaN
IWL	A lot NaN
SBW	A lot NaN
SBL	A lot NaN
B&WW	A lot NaN
B&WL	A lot NaN

3.1.2 Players

Field	Reason for Dropping
first_initial	No useful info
player_url	No useful info
height_ft	have height_cm
height_inches	height_cm
birth_day	Have date
birth_month	Have date
birthplace	No useful info
weight_lbs	Have weight_kg
handedness	No useful info
residence	No useful info
backhand	No useful info
first_name	Have full_name
last_name	Have full_name

3.2 Data Cleaning

- During the data cleaning process, some of the numerical columns in players.csv were processed to remove percent and comma symbols and convert them to proper types. This step was necessary to ensure consistency in the data and make it more suitable for analysis. For instance, values such as "19%" were converted to 0.19, while "6,66" was converted to 6.66. This conversion will enable more accurate calculations and comparisons while performing data analysis.

- To fill in missing values in the "turned_pro" column, we take the average of the differences between "turned_pro" and "birth_date", and add this value to "birth_date".

- There were many missing values in some columns. We dropped certain columns and replaced the missing values in others with their respective average values.

3.3 Data Construction

- To determine the country of each match, we utilized the Yandex Geocoder API. By leveraging the city where the match took place, such as Sydney, we were able to add the corresponding country (Australia) to the dataset. The "players.csv" file contained a field for the player's country, represented by a three-letter abbreviation code (e.g., RUS, USA). To convert these codes into their full names, we utilized Wikipedia's mapping table. However, it was necessary to manually update some names to ensure they matched the country locations of the matches. For example, if a player's country was listed as Great Britain but the match took place in the United Kingdom, we adjusted the player's country accordingly.

- In order to enrich the players table with additional features, we developed a Python script. This script parsed player characteristics from the official website of Men's Professional Tennis, resulting in the addition of 18 new features to the dataset.

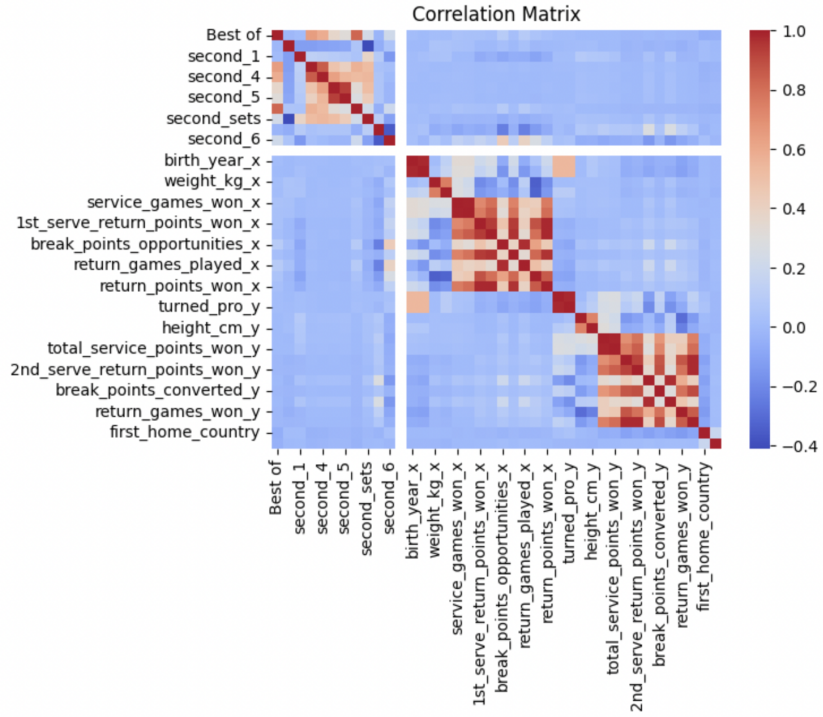


Figure 4: Correlation between features

3.4 Data Integration

- In order to expand the player's characteristics in the "matches.csv" file, we performed a join operation with the cleaned "players.csv" file. However, due to the previously mentioned issue, the joining field had to be the full name of the players. Additionally, inconsistencies between the two tables resulted in some matches being unable to retrieve player details. As a result, approximately 11,000 matches were excluded from further analysis. Nevertheless, we successfully incorporated certain features that could prove useful during the modeling stage.

3.5 Data Formatting

- At this stage, the joined table consisted of rows with both a winner and a loser. To account for this, we made the decision to duplicate all the rows but with the players swapped. Essentially, each match occurred twice, with the winner of the match listed as the first player in one instance and as the second player in the other instance.

- Before proceeding with the modeling phase, the data was shuffled to ensure randomness and eliminate any potential biases.

3.6 Data quality

- In addition to Tableu, our data science team utilized Python with Pandas and Numpy to analyze the missing or poor-quality data. Here are the findings from our investigations:

- One major issue with the "matches.csv" file is the absence of player IDs, making it challenging to identify the participants solely based on their names.

- Initially, a significant portion of the players in "players.csv" had missing or zero values for most of their attributes (approximately 80%). However, since most of these players were not present in the "matches.csv" file, the impact on the results was not significant.

- We encountered cases where multiple players in "players.csv" shared identical full names, making it impossible to determine which of them participated in specific matches. This issue proved to be quite frustrating, and a potential improvement for the project would be to acquire more reliable match data.

- Additionally, there were 16,208 matches without winner's points and 16,288 matches without loser's points, indicating missing information in these fields.

- Furthermore, there were 21 matches without the winner's rank and 129 matches without the loser's rank.

- Lastly, around 100 players had missing information such as name, height, or weight, further contributing to data incompleteness.

4 Modeling

4.1 What is TPOT?

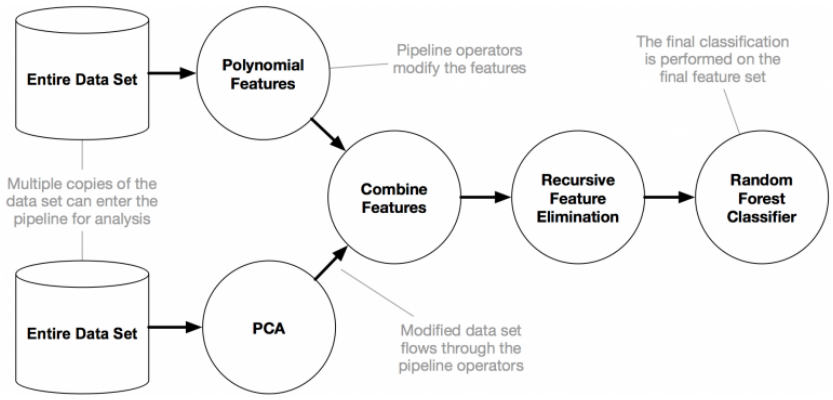


Figure 5: TPOT

TPOT, short for Tree-based Pipeline Optimization Tool, is an automated machine learning tool developed by the Data Science and Automation team at DataRobot. It uses genetic programming to optimize a pipeline of machine learning models, feature preprocessors, and data transformations to find the best model for a given dataset.

With TPOT, users only need to provide their dataset and specify the target variable, and TPOT takes care of the rest. It evaluates a wide range of algorithms and preprocessors, including feature selection, dimensionality reduction, and scaling, to identify the most effective combination for the given dataset.

TPOT uses a genetic algorithm to search through the space of possible pipelines. It starts with an initial population of randomly generated pipelines and evolves them through successive generations, with each generation consisting of a new population of pipelines generated through genetic operations such as mutation, crossover, and selection. TPOT uses a fitness function to evaluate each pipeline, which is usually the cross-validated accuracy score of

the pipeline on the training data.

One of the benefits of TPOT is that it can handle a wide range of data types, including numerical, categorical, and text data. It can also handle missing values and automatically preprocess the data before training the model.

In addition to its automation capabilities, TPOT also allows for manual intervention in the pipeline optimization process. Users can specify the range of algorithms and preprocessors to be considered in the optimization process and can also include custom transformers and models.

To use TPOT, you will need to install it via pip or conda, and then import it into your Python script or notebook. You can then create an instance of the TPOT class and call its fit method to run the pipeline optimization process. Once the optimization process is complete, you can call the TPOT instance's score method to evaluate the best pipeline on a held-out test set.

In conclusion, TPOT is a powerful tool for automating the machine learning pipeline optimization process. Its ability to handle a wide range of data types and missing values, along with its flexibility to allow for manual intervention, make it a valuable tool for data scientists and machine learning practitioners. With TPOT, you can save time and effort in the pipeline optimization process and focus on other aspects of your machine learning project.

To ensure that we used the best modeling technique with optimal parameters, we employed TPOT.

Before applying TPOT, we applied PCA to the data to reduce the number of columns and avoid overfitting. To determine the optimal number of components for PCA, we utilized the Elbow Method with PCA. The Elbow Method is a technique used to identify the optimal number of components by plotting the explained variance against the number of components and identifying the "elbow point" on the graph. Although PCA could reduce the accuracy a little bit, it did. But it is better than overfitting on the old data, and whatever new data the user want to predict, it will give wrong predictions.

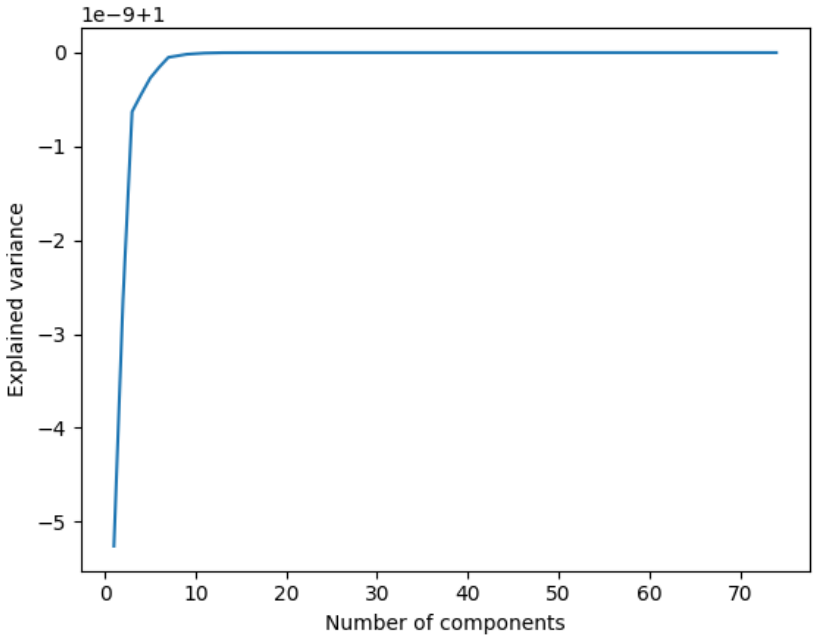


Figure 6: Components - Variance

After identifying the optimal number of components for PCA (e.g., 10 components), we split our dataset into training and testing sets using an 80:20 ratio. To address the lengthy fitting process in the TPOTClassifier, we employed the max_time_mins parameter set to 60, limiting the model's runtime. Without applying PCA, the model typically takes twice as long to run, considering 20 populations and 5 generations in TPOT.

The TPOTClassifier yielded the best model, as determined through the evaluation process.

To further evaluate and improve our model, we created another model that employs the same dataset. This additional model utilizes cross-validation to identify the optimal parameters. By combining the predictions of both models, we can enhance the credibility of our betting profits and assess TPOT's performance regarding date dimensionality. For this second model, we chose to use Random Forest.

5 Evaluation

In order to evaluate the performance of our two models, we utilize the "roc_auc_score" metric, which is commonly employed for assessing binary classification models. This metric calculates the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings.

The primary objective of our models is to provide an efficient tool for doubling the invested money through a minimum number of consecutive bets. To achieve this, we calculate the betting coefficient using the following equation:

$$coef = \frac{1}{prob + prof * 0.5}$$

where:

- coef: is the calculated betting coefficient.
- prob: is the probability predicted by our model.
- prof: is the profit percentage the bookmaker usually keeps to himself.

The Money change function is written as follows: If the bet is won:

$$MAB = MbB + (MbB \times rf \times (prob - 0.5) \times coef)$$

If the bet is lost:

$$MAB = MbB - (MbB \times rf \times (prob - 0.5))$$

where:

- coef: is the calculated betting coefficient.
- prob: is the probability predicted by our model.

- MAB: is the balance after the bet is settled.
- MbB: is the balance before the bet is settled.
- rf: is the risk factor.

Note that the MbB represents all the money owned before a bet. Based on the client's desire of the trade-off between risk and safety, he can choose a value for the risk factor in the interval (0, 1), based on which he will gamble with the amount $MbB \times rf \times (prob - 0.5)$.

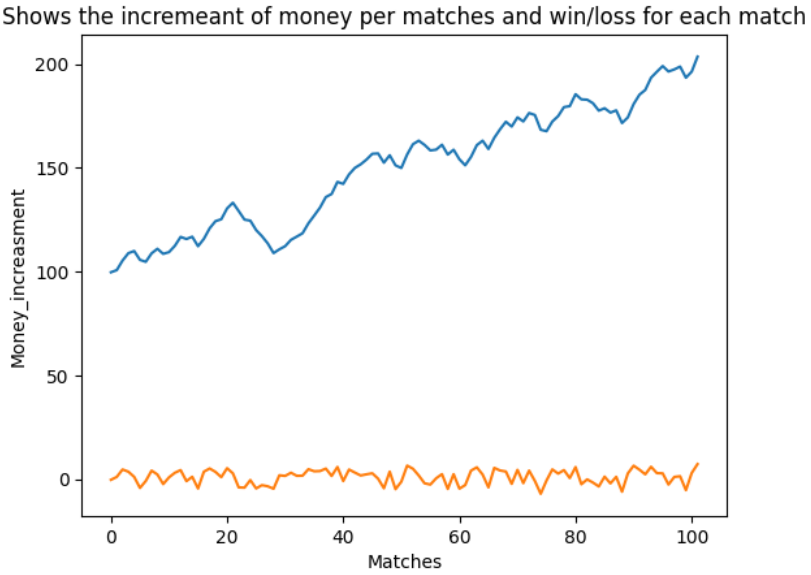


Figure 7: Matches - Profit

By applying this approach, we aim to increase profits on a sorted test dataset based on the timestamp. We visualize the betting profits and losses for each match bet. The graph above illustrates the change in the initial money (represented by the blue line) and the wins and losses (represented by the orange line) after each bet. The significant increase in the initial money indicates the effectiveness of our model.
To expedite the process of achieving the target money, we can increase the Risk Factor. However, it is important to note that a very high Risk Factor can lead to bankruptcy. The graph below demonstrates the relationship between the Risk Factor and the number of matches required to double the

initial money. We observe that the optimal value for Risk Factor is 0.1, which allows for doubling the money in 100 bets on matches.

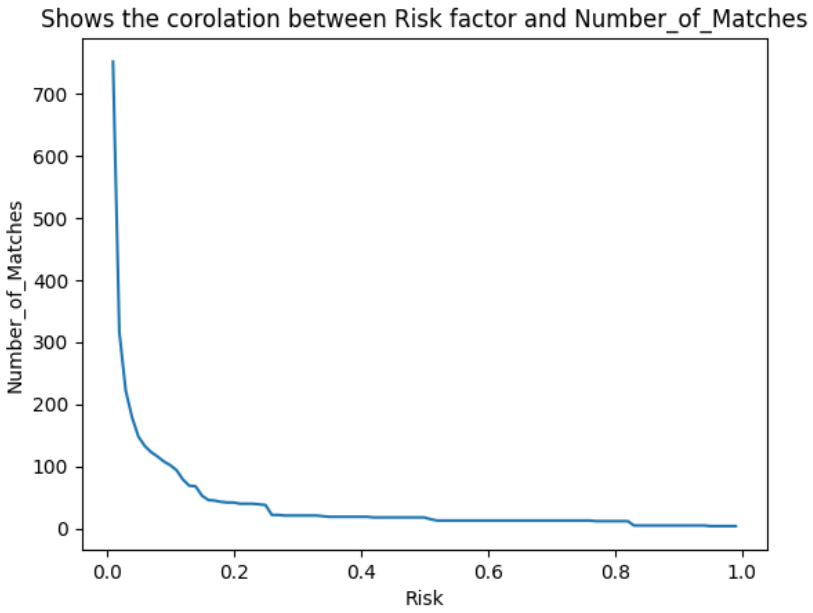


Figure 8: Matches - Profit

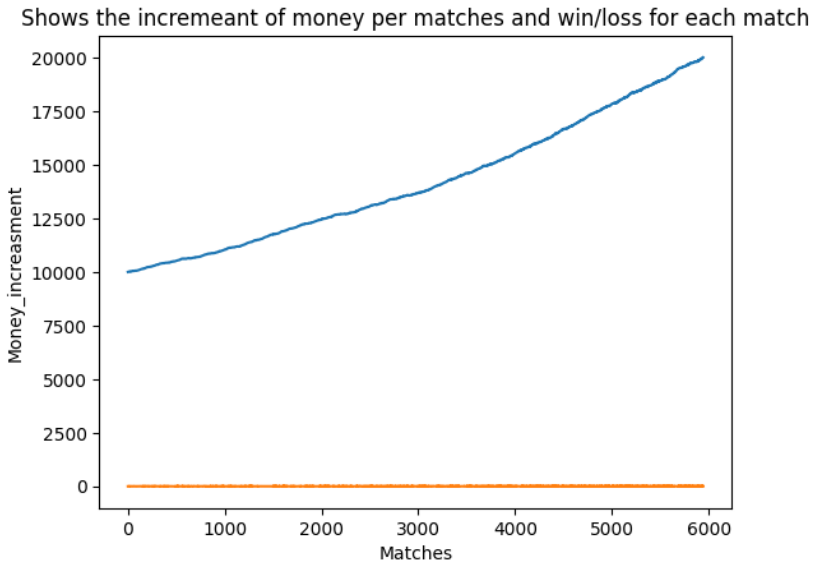
On the other hand, selecting a very low Risk Factor results in the graph shown below. It illustrates the lower monetary gain/loss indicated by the orange line. However, with 6000 bets on matches, we can safely double the money.

Overall, our evaluation demonstrates the effectiveness of our models in achieving the desired profit by employing appropriate Risk Factors and maximizing the winning predictions.

6 Deployment

6.1 Deployment Plan

- We have developed a Telegram bot that assists our customers in earning money by predicting the results of upcoming matches. Our team



utilized several technologies, including Python 3.9, aiogram, Docker, and dashboard.render.com to deploy our bot to an outsourced location.

- The use of Python 3.9 allowed us to implement advanced machine learning algorithms that can effectively analyze data on past matches, player performance, and other key metrics to predict future outcomes accurately. We also utilized aiogram, a Python asynchronous library, to develop the bot's functionality and ensure seamless integration with Telegram.

- Docker, an open-source containerization platform, allowed us to package the bot and its dependencies into a single, portable unit. This approach made it easy to deploy the bot to any environment, including a cloud-based platform.

- Finally, we used dashboard.render.com, a cloud-based platform, to host our bot and manage its deployment, updates, and scalability. This platform provided us with a reliable and secure infrastructure to ensure that our bot runs smoothly and efficiently.

- Overall, our team leveraged various technologies to create a powerful

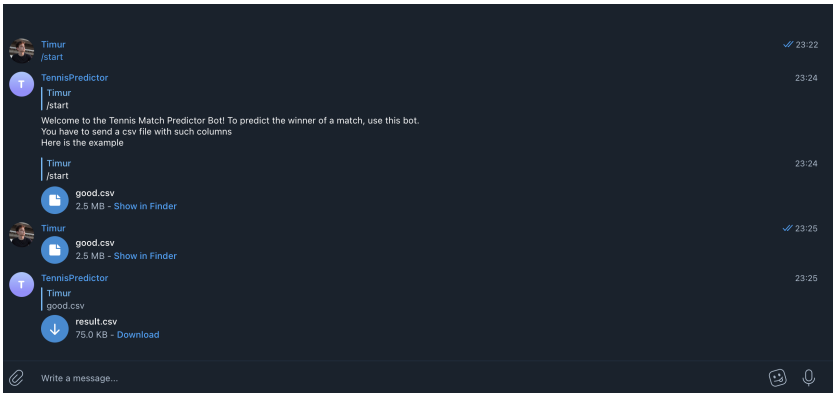
and reliable Telegram bot that can help our customers earn money by predicting match results. We are confident that our bot will provide valuable insights to our customers and help them make informed decisions when placing bets on upcoming matches.

6.2 Monitoring and Maintenance Plan

- Our current hosting service is providing us with limited resources, specifically only 0.1 CPU and 1GB of memory. These limitations may be hindering our ability to scale our business and meet the evolving needs of our customers. As a result, we believe it is important to explore alternative hosting options that can accommodate our growing business needs.

- In addition to finding a hosting service that can provide more resources, we also recognize the importance of engaging with our customers and expanding our horizons. By understanding the needs and preferences of our customers, we can tailor our products and services to better meet their needs and ultimately drive growth for our business.

- Overall, we see the need to address our hosting limitations and expand our customer outreach as key steps in scaling our business and achieving long-term success.



Builds too slow? Upgrade to a paid instance type to go faster. Learn more about free instance type limits.

May 17, 2023 at 11:23 PM In progress

b55029f pre final

Search logs Search Maximize Scroll to top

```
May 17 11:25:52 PM  #9 DONE 30.7s
May 17 11:25:52 PM
May 17 11:25:52 PM  #10 exporting to docker image format
May 17 11:25:52 PM  #10 exporting layers
May 17 11:26:27 PM  #10 exporting layers 35.1s done
May 17 11:26:27 PM  #10 exporting manifest sha256:f296675411a320803a15b9a6c4dc1366a74042ae8dac7bd32ad7b48055178f0b done
May 17 11:26:27 PM  #10 exporting config sha256:bbccbf4e64c7b3b7f1a7ac4463d091a00bad79f4d8df74e450e7648e89e8a8ad done
May 17 11:26:48 PM  #10 DONE 55.1s
May 17 11:26:48 PM
May 17 11:26:48 PM  #11 exporting content cache
May 17 11:26:48 PM  #11 preparing build cache for export
May 17 11:27:06 PM  #11 DONE 18.2s
```

Usage

Total usage: **1.37 hours** (month-to-date)

Average usage: **1.37 hours/day**

You have used **1%** of your monthly quota across all of your free instances. View breakdown