George Smith

Final Portfolio Draft

Completing my M.S. in Applied Data Science has been essential for my development as a data professional. I have been introduced to essential tools and coding languages including Python, R, SQL, and AWS. Three projects that exemplify my newly developed data analysis skills include projects that I created in my Scripting for Data Analysis (IST 652), Database Administration & MGMT (IST 659), and my Quantitative Reasoning for Data Science (IST 772) courses.

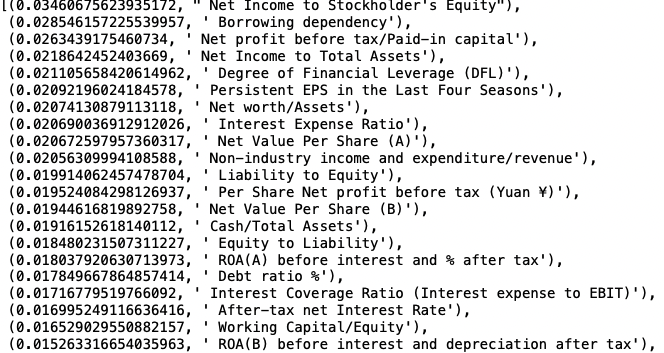
In my scripting for data analysis class as a final presentation I leveraged web scraping, machine learning, and data cleaning techniques python techniques to answer three specific questions related to bankruptcy projections. These included:

1. Are there any companies in the S&P 500 with financial similarities to bankrupt companies?
2. What Financial indicators heavily influence if a company is likely to face bankruptcy?
3. What financial indicators have little influence on the likelihood of a company facing bankruptcy?

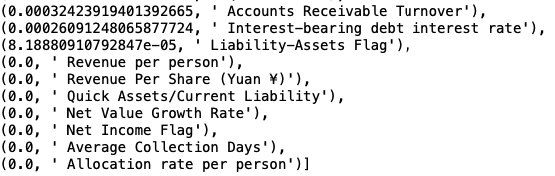
The final metrics I considered in this analysis were ROA, Current Ratio, Quick Ratio, Debt/Equity, and Gross Margin. To develop this analysis two datasets were considered. The first dataset was provided by the Taiwan Economic Journal. The dataset included 95 columns and 6,820 rows and all columns represented various financial metrics. A Boolean indicator was included in this dataset that represented if a company had gone bankrupt. The second dataset included web scraped financial information for every ticker represented in the S&P 500. I scraped this information from FinViz.com and included 18 columns by 505 rows of various financial metrics. The analysis I conducted on Operating Gross Margin showed that the average Operating Gross Margin for bankrupt companies was 60%. I found that many companies in the S&P 500 had similar operating gross margins including Brown-Forman Corporation, Colgate-Palmolive Company, Equifax, Hasbro, Coca-Cola Co, Mid-American Apartment Communities, Health peak Properties, Teradyne Inc., Take-Two Interactive Software, WEC Energy Group Inc. 13 total companies had values within 1% of the average Operating Gross Margin of bankrupt companies. The average ROA of bankrupt companies was 50%. eBay was the only company with a similar ROA of 51.5%. The average current ratio of bankrupt companies was 416,729. No companies in the S&P 500 had current ratios that were this high. A maximum value of 2,750,000,000 was found in the dataset representing bankrupt companies. Because of this outlier it was hard to draw any conclusions about the Current Ratio. The average Quick Ratio of bankrupt companies was 7,257,160. The largest Quick Ratio found in the S&P 500 was Aadi Bioscience with a value of 14. Due to the different values, it was hard to draw any conclusions about Quick Ratio. The average Liability to Equity Ratio was 28%. 10 companies in the S&P 500 had similar Liability to Equity ratios within 1% of bankrupt companies. These include Chubb Ltd, Coterra Energy Inc, Harford Financial Services Group Inc., Metlife Inc, Mohawk Industries, Inc, New Corp Class A, Paychex Inc, Principal Financial Group, Progressive Corp, Exxon Mobil Corp.

In the second part of this analysis, I used random forest modeling considering all 95 of the financial metrics found in the Bankruptcy data to determine what financial metrics are most and least influential for calculating bankruptcy. I found that Net Income to Stockholders Equity, Borrowing Dependency, Net Profit Before Tax/Paid-in Capital, Net Income to Total Assets, and Degree of Financial Leverage were the most influential financial metrics. The least influential financial metrics were Allocation Rate Per Person, Average Collection Days, Net Income Flag, Net Value Growth Rate, Quick Assets/Current Liability, Revenue Per Share, and Revenue Per Person.

**Figure 1** depicts the financial metrics that were most influential for predicting bankruptcy



**Figure 2** depicts the financial metrics that were least influential for predicting bankruptcy



What I learned from this project is that predicting bankruptcy using only general financial metrics is very difficult. To better make these predictions I think it would be essential to leverage regulatory notices related to corporations or individuals associated with specific corporations. I believe these signals would be essential in determining if a company is likely to face bankruptcy.

In my Database Administration & MGMT (IST 659) class I leveraged SQL, R, Microsoft Access, and database modeling techniques. I created a fictitious database implementation project for a soccer betting website. The name of the website was GOOOAL Bets. GOOOAL Bets is a premier soccer betting website known for its precise lines and exotic prop bets. GOOOAL Bets also offers analytics services for entitled users. To have an edge over the competition, GOOOAL Bets is looking to implement its own in-house database. Implementing this database will allow them to own an almanac of statistical resources leading to more precise and insightful betting opportunities for their customers.

The stakeholders included in this database implementation project include the CEO, board of directors, all employees, and any customers who are either placing bets or using the analytical services. The CEO and board of directors are hoping that this project will lead to increased profits due to the precise lines and added variety of bets the database will allow them to offer. Employees will have an easier time creating betting opportunities for customers due to the easily accessible data. Customers will enjoy a wider variety of bets, and more accurate analytical services.

The following business rules were essential for the database implementation

* A player can only play one position
* A player can play for multiple teams during a season
* A team has many players
* A team can only have one coach
* A coach can only coach one team
* A team can only be from one city
* A team can only have one record
* Two teams play per match

Developing a business glossary was essential for understanding the database terminology

**Player -** An individual taking part in a soccer match.

**Team -** A group of Players forming one side in a soccer match.

**Position-** The joint arrangement of a Team on its field of play during a game and to the standardized place of any individual Player in that arrangement.

**Coach-** A person who teaches or trains a Player.

**City-** A location where a Team is from.

**Match-** A contest in which Teams compete against each other.

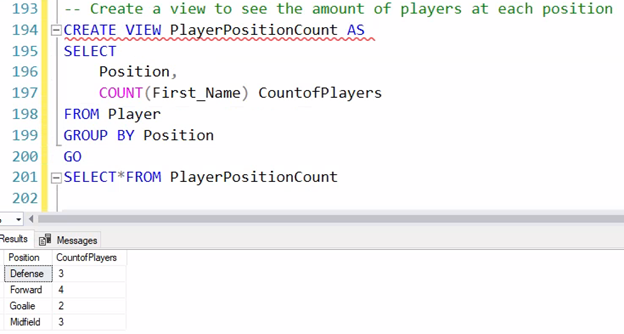
**Captain-** The player who leads the Team.

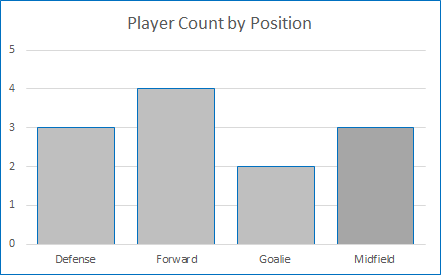
The main purpose of the database was to answer the following business questions using SQL code

Q: What is the total number of players by position?

A: Defense 3, Forward 4, Goalie, 2 , Midfield 3

**Figure 3** depicts the SQL code used to answer the above data question

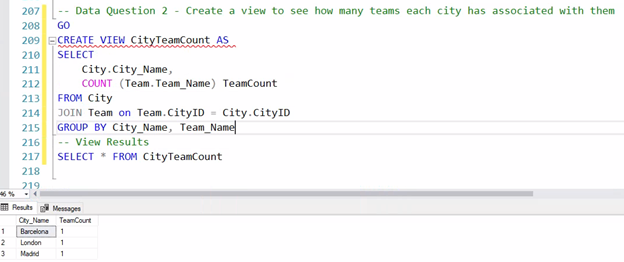


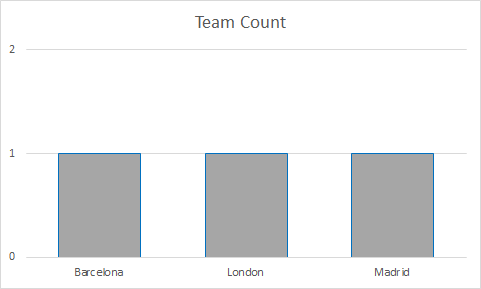


Q: How many teams does each city have associated with them?

A: Each City has 1 team associated with them

**Figure 4** depicts the SQL code used to answer the above data question

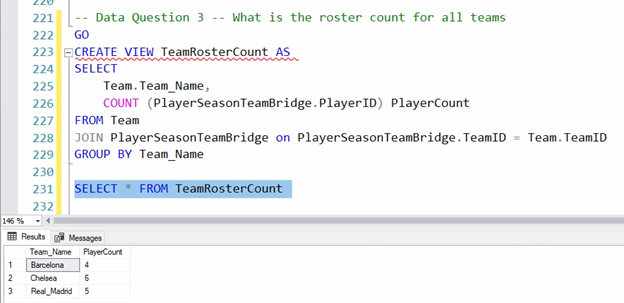


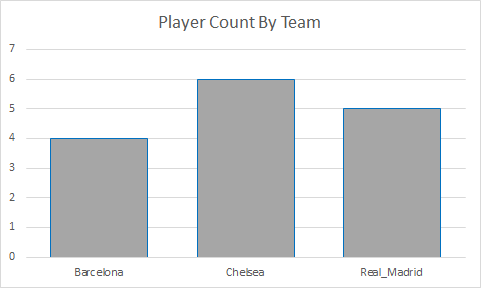


Q: What is the roster Count for all teams?

A: Barcelona 4, Chelsea 6, Real\_Madrid 5

**Figure 5** depicts the SQL code used to answer the above data question

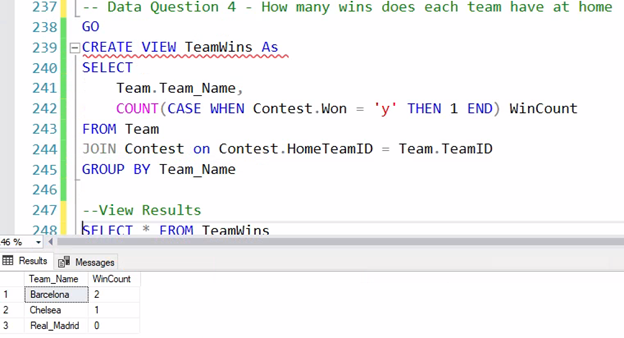


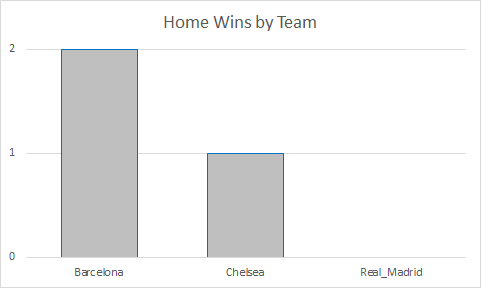


Q: How many wins does each team have at home

A: Barcelona 2, Chelsea 1, Real\_Madrid 0

**Figure 6** depicts the SQL code used to answer the above data question

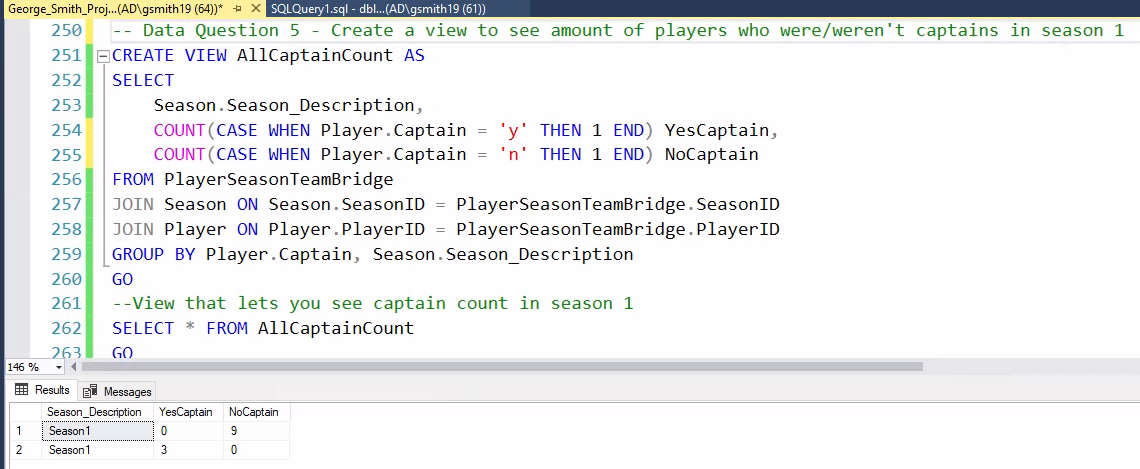


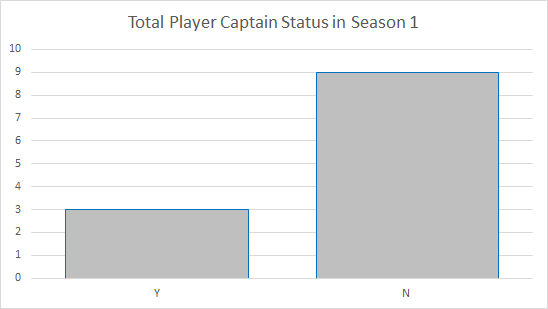


Q: How many players were captains in season 1 versus players that were not captains in season 1

A: 3 players were captains in season 1, 9 players were not captains in season 1

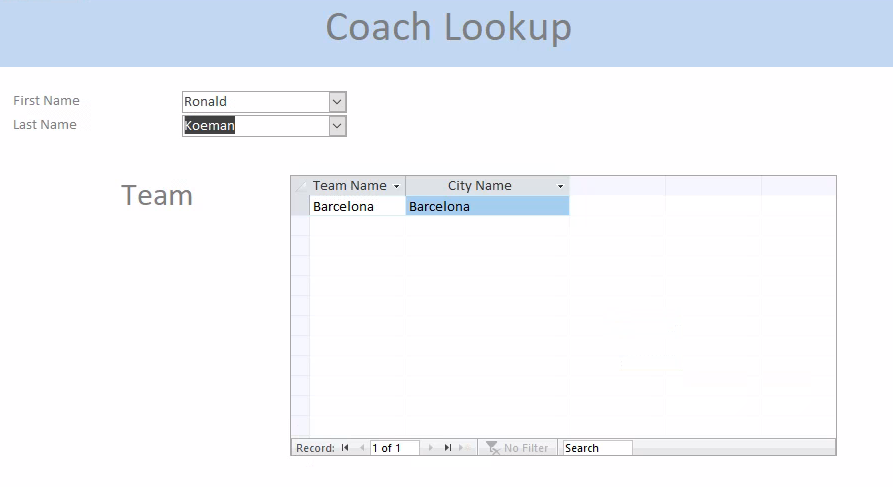
**Figure 7** depicts the SQL code used to answer the above data question





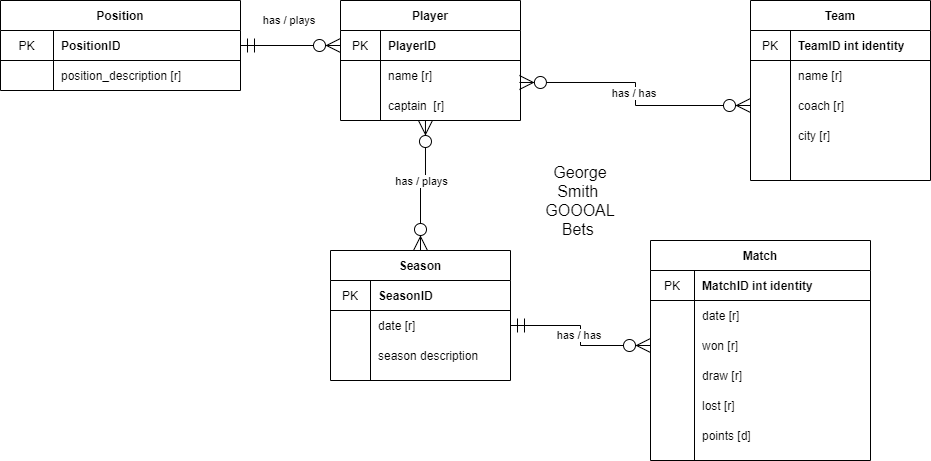
In addition to creating a database and filling it with data I created a report using a Microsoft Access that could look up a team’s name and associated city by selecting a coach’s first or last name.

**Figure 8** depicts the Coach lookup report in Microsoft Access

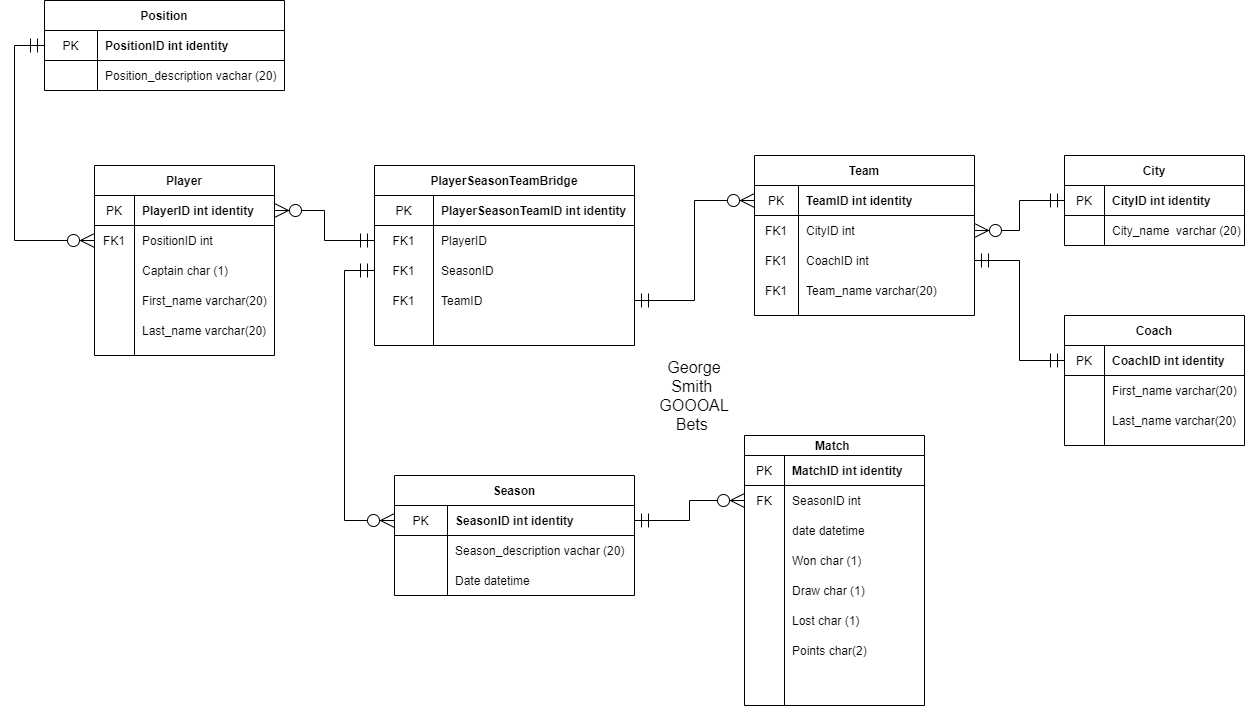


In order to successfully answer my data questions, it was essential to have accurate conceptual and logical models

**Figure 9** depicts my conceptual model

****

**Figure 10** depicts my logical model



What I learned from this project is that a good logical and conceptual model are necessary for creating a database. This is something I continued to refine throughout the project. I was able to create views while using select statements to answer all my data questions. I created bar charts using R to generate bar data visualizations for each of my data questions. I created a form in access that can look up a coach’s team and the associated city by selecting either their first or last name.

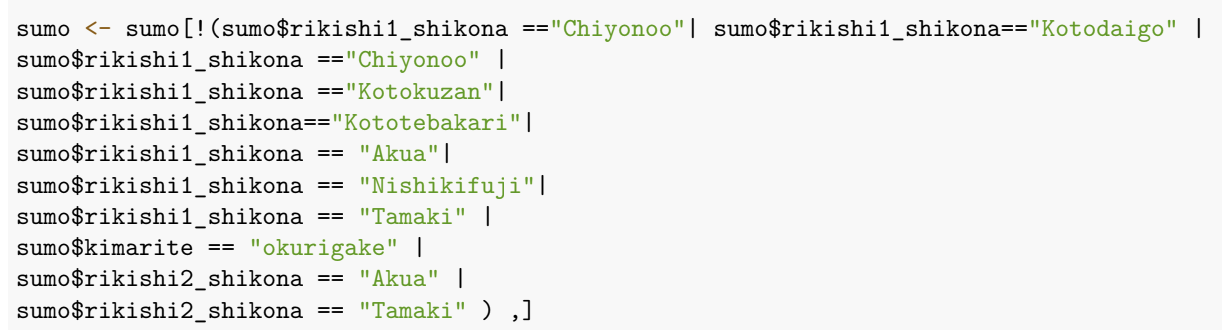
At the start of this class, I had only heard of the terms logical model, conceptual model, and SQL. My first assumption was that these models look simple enough to create and understand. I quickly learned that these diagrams require tons of planning and can be extremely difficult to create. I was very intimidated when it came time to write code in SQL, as I have minimal coding experience. I found this portion of the project to be very enjoyable, and I feel that the SQL syntax is very logical.

In my Quantitative Reasoning for Data Science (IST 772) class I used sumo wrestling data to predict potential winners of sumo matchups. The data is an aggregation of results of every bout (top two divisions) in Japanese sumo wrestling grand tournaments (honbasho) from 1985 until 2019 scraped from <http://sumodb.sumogames.de/Results.aspx?b=198301&d=15&simple=on>. This data set required a ton of cleaning to be model ready. The first step I took was to remove any columns that designated a win, as these columns would not be usable for a predictive model. I used the dplyr package to drop the rikishi2\_win, rikishi1\_result, and rikishi2\_result column because of their relationship to rikishi1\_win. Additionally, I dropped the below records as they had little representation in the data and would only skew my model.

**Figure 11** depicts the sumo ranks that were dropped from my data due to underrepresentation

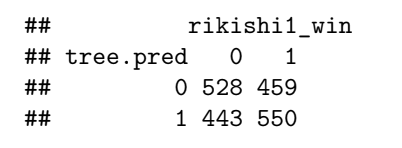


**Figure 12** depicts the sumo wrestlers that were dropped from my data due to underrepresentation



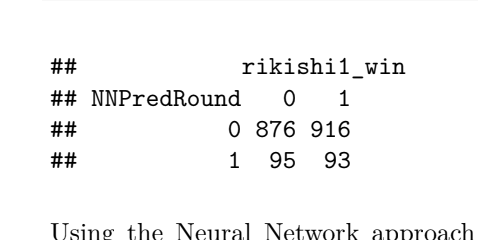
Once the data was clean, I tested and trained the data to predict winners of a sumo match using the decision tree modeling technique. Using the decision tree approach, I was able to predict 528 instances of losses and 550 instances of wins correctly. This means that the decision tree model was able to predict sumo wrestling losses with 54%accuracy and wins with 56 % accuracy. Overall, I was satisfied with this result as my model was able to have better results than simply guessing. Since I knew absolutely nothing about sumo wrestling this model would be usable for my purposes.

**Figure 13** depicts the win loss predictions of my decision tree model



I decided to see how a neural network model would fair for creating a predictive model with this dataset. To run a neural network model, I needed to transform my categorical data. I used the fast dummies package to do this so that I could include all my data in this model. Unfortunately, this model did not yield useable results. Using the Neural Network approach, I was able to correctly predict 876 losses correctly and 93 wins correctly. 90% of the loss predictions were correct, compared to only 9% of the win predictions being correct. I believe these poor results are due to having to create dummy variables for most of the data set so that the data would be usable for the Neural Network Model. I believe models such as the decision tree that can use categorical data would be better suited to make predictions on the Sumo data set.

**Figure 14** depicts the win loss predictions of my neural network model



Using machine learning techniques including Decision Tree and Neural Network Models I was able to successfully predict if a sumo wrestler would win or lose a given match. Based on the above analysis it appears that the Decision Tree model was better suited for the Sumo data used as the Neural Network was able to only predict 9% of wins correctly. Although 56%-win prediction accuracy is not incredible it is an improvement over simply guessing the result of a given match.