Final Report:

Online Texas Hold'em Poker strategy analysis

Problem statement

Texas Hold'em poker is among the most popular card game for almost two centuries. For majority of the game's lifetime people have been playing it by feel, it was not until recent decades when people started to explore the science behind the game. It turned out that poker was much more than a game of chance, where proper strategies can be implemented to make one a consistent winner.

However, winning poker strategies are very complex, they involve odds, combinatorics, observation, and self-control, and professional poker players spend years to perfect their strategies. Recreational players enjoy playing poker, but do not have time nor the dedication to develop and practice a winning strategy. The goal of this analysis is to find a simple strategy that players can pick up and use immediately in the game. Players who adapt this simple strategy are unlikely to become winning players, but with it they should be able to enjoy the game while losing less money than if they had no strategy at all.

The benefactors of this analysis would be recreational poker players, and poker room owners or online poker site. Recreational poker players for obvious reasons, they will have fun at the game while "paying less fee". If recreational players, which make up over 95% of the poker population, are able to stay in the game longer, poker room owners and online poker sites will also make more profit.

Data Wrangling

Online poker sites usually have records of the hands that are played on their site. The data for this analysis were records of all poker hands played on an online poker site from 07/01/2009 to 07/23/2009.

The records range from $0.50/$1.00 blinds to $5.00/$10.00 blinds stakes. The lowest stake ($0.50/$1.00 blinds) poker hands were chosen, because recreational players are most likely to play at low stakes.

324,011 poker hands were chosen, and over 8000 players were involved in these hands. All the poker hands were recorded across 325 text files, with each text file contained 1,000 poker hands in descriptive format. The following figure shows a snapshot of a poker hand in descriptive format.

```
Stage #3017235188: Holdem  No Limit $1 - 2009-07-01 00:00:10 (ET)
Table: ALABAMA ST (Real Money) Seat #4 is the dealer
Seat 4 - s32h30cC3rPhG5FiSCU42g ($55.50 in chips)
Seat 5 - uwsIGbIB4ktlhz44JKQKDQ ($247.10 in chips)
Seat 1 - ahN79dgDL8C99nHik5Up7Q ($14.60 in chips)
Seat 2 - 9tEZRm6oI+e1Tz0x72b0vQ ($200 in chips)
Seat 3 - AiiJXMM0CfYl69+Nq3jyfA ($96.50 in chips)
uwsIGbIB4ktlhz44JKQKDQ - Posts small blind $0.50
ahN79dgDL8C99nHik5Up7Q - Posts big blind $1
9tEZRm6oI+e1Tz0x72b0vQ - Posts $1
*** POCKET CARDS ***
9tEZRm6oI+e1Tz0x72b0vQ - Checks
AiiJXMM0CfYl69+Nq3jyfA - Folds
s32h30cC3rPhG5FiSCU42g - Folds
uwsIGbIB4ktlhz44JKQKDQ - Calls $0.50
ahN79dgDL8C99nHik5Up7Q - Checks
*** FLOP *** [9s 7h 2c]
uwsIGbIB4ktlhz44JKQKDQ - Checks
ahN79dgDL8C99nHik5Up7Q - Checks
9tEZRm6oI+e1Tz0x72b0vQ - Checks
*** TURN *** [9s 7h 2c] [9d]
uwsIGbIB4ktlhz44JKQKDQ - Checks
ahN79dgDL8C99nHik5Up7Q - Checks
9tEZRm6oI+e1Tz0x72b0vQ - Checks
*** RIVER *** [9s 7h 2c 9d] [9c]
uwsIGbIB4ktlhz44JKQKDQ - Checks
ahN79dgDL8C99nHik5Up7Q - Checks
9tEZRm6oI+e1Tz0x72b0vQ - Checks
*** SHOW DOWN ***
uwsIGbIB4ktlhz44JKQKDQ - Shows [8d As] (Three of a kind, nines)
ahN79dgDL8C99nHik5Up7Q - Mucks
9tEZRm6oI+e1Tz0x72b0vQ - Mucks
uwsIGbIB4ktlhz44JKQKDQ Collects $2.85 from main pot
*** SUMMARY ***
Total Pot($3) | Rake ($0.15)
Board [9s 7h 2c 9d 9c]
Seat 1: ahN79dgDL8C99nHik5Up7Q (big blind) HI: [Mucked] [5c 6c]
Seat 2: 9tEZRm6oI+e1Tz0x72b0vQ HI: [Mucked] [5s Jc]
Seat 3: AiiJXMM0CfYl69+Nq3jyfA Folded on the POCKET CARDS
Seat 4: s32h30cC3rPhG5FiSCU42g (dealer) Folded on the POCKET CARDS
Seat 5: uwsIGbIB4ktlhz44JKQKDQ (small blind) won Total ($2.85) HI:($2.85) with Three of a kind, nines(ace kicker) [8d As - B:9s,B:9d,B:9c,P:As,P:8d]
```

Figure 1: snapshot of a poker hand in description format

Python's REGEX module was used extensively to extract information from these text files. Some of the important features of the poker hands are: hand ID, player ID, stack size, player position, action and amount of each street etc.

After extracting information with REGEX, the features of interest were temporarily stored in lists. Then python's PANDAS module was used to compile all the lists into a data frame. The data frame was set up so each row indicated the features of a player in a particular hand. The resulting data frame that displayed every player in every poker hand had 1,767,588 rows, and 16 columns. Figure 2 shows a part of this data frame.

hand_id	player_id	seat	stack	position	post	preflop	p_amount	flop	f_amount	turn	t_amount	river	r_amount	player_num
3017237436	vETYfpoA+FhBercnDPJrRw	5	197	dealer	0	Folds	0	NA	0	NA	0	NA	0	4
3017237436	DeZAZcPNNQ5w+Wb+5ujZdA	6	200.30	small blind	0.5	Raises	2.5	NA	0	NA	0	NA	0	4
3017237436	AiiJXMM0CfYl69+Nq3jyfA	2	78.50	other	0	NA	0	NA	0	NA	0	NA	0	4
3017237436	id+sbECX+Ydl8qhMhpje+g	3	81.60	big blind	1	Folds	0	NA	0	NA	0	NA	0	4
3017235188	s32h30cC3rPhG5FiSCU42g	4	55.50	dealer	0	Folds	0	NA	0	NA	0	NA	0	5

Figure 2: example of the data frame displaying player details

Exploratory data analysis

Throughout the exploratory data analysis process, the data frame created in the data wrangling section was transformed into rows grouped by players. This was done because the final goal was to find the players who are doing well, then identify the attributes of those players. In this transformation, many player specific features were created. These features included: number of hands played, net amount won, VPIP (poker term, indicating the player's willingness to get involved in pots), average pot size etc. Figure 3 shows a section of the early stage of this data frame.

	hands_played	pots_won	amount_won	to_inv	net_win	per_hand	vpip_count	inv_count	vpip	% won	avg_p_size
hc0LUofSVtJkKIO0r20FzA	2147	199.0	4536.24	3919.26	616.98	0.287368	359	770	0.167210	0.258442	22.795176
gxztVvz8QgeSaABFE8/xYQ	2219	162.0	2950.60	2511.25	439.35	0.197995	311	748	0.140153	0.216578	18.213580
gm8F7k++eftjEjB5FWxUTA	1671	239.0	5397.90	5041.34	356.56	0.213381	356	816	0.213046	0.292892	22.585356
lr5ondonH45KZUPlsjjOZg	1667	176.0	2910.37	2575.30	335.07	0.201002	407	705	0.244151	0.249645	16.536193
sBCULyaFD9K2rD/+eGv7Eg	1522	150.0	2724.30	2501.24	223.06	0.146557	302	588	0.198423	0.255102	18.162000

Figure 3: section of data frame grouped by players

The data frame in Figure 3 was indexed by player ID. Figure 4 shows a graph of net win (and loss) by all the players.

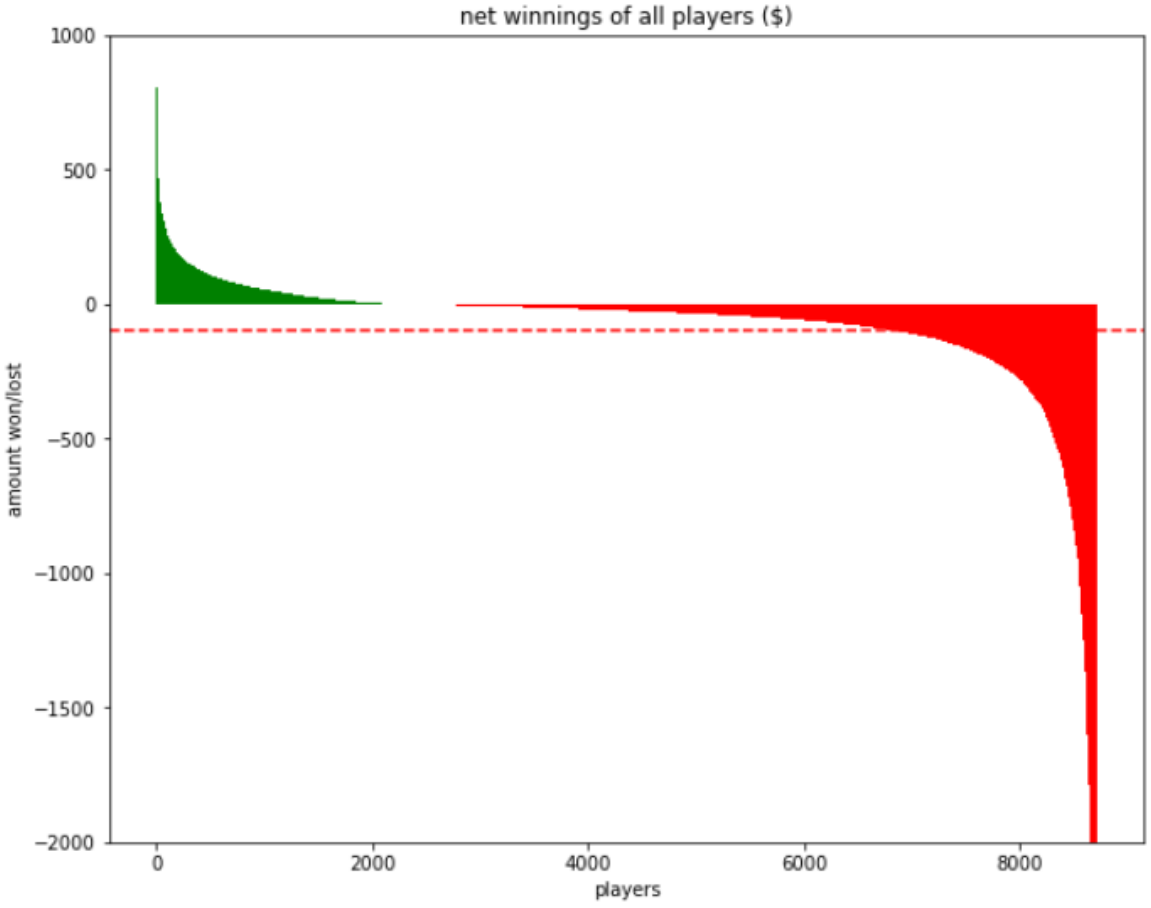

Figure 4: net win and loss of all players

Figure 4 showed that there were over 8000 unique players involved, which matched the number of rows of the data frame grouped by players. It also showed that there were many more losing players than winning players. This was due to the casino rake which are taken from every pot, which reduced the sum of net win and loss of all players below $0 (which should be the case in a zero-sum game like poker). In fact, the breakeven point, which was indicated by the red horizontal dotted line in the graph, was observed to be around -$80 per player. Which meant that the casino took over $640,000 in total rake across all the players. Finally, it was observed average losing players lost much more than average winning players won, this was partly due to the enormous amount of rake, also partly due to players tendency to chase losses and stop after wins.

Many more player features were added to the data frame. Such as aggression on each street, number of players per hand, winning per hand played etc. Most of these features are derived from pre-existing features, which may or may not be important features that defined a good poker player. Only players who played 1000 or more hands of poker were selected, to limit the effect of random luck. All the features were graphed to observe possible relation with either net win or winning per hand, some of the most correlated features are shown in the graphs below.

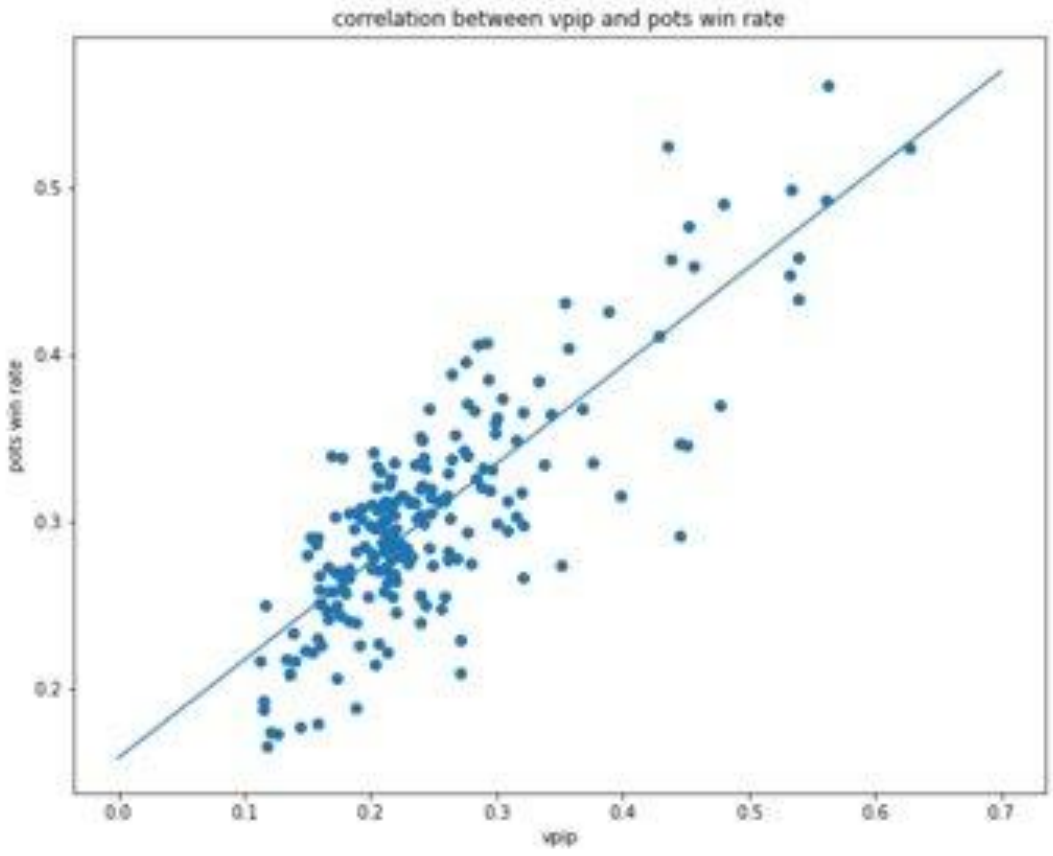

Figure 5: effect of VPIP (player's willingness to get involved in pots)

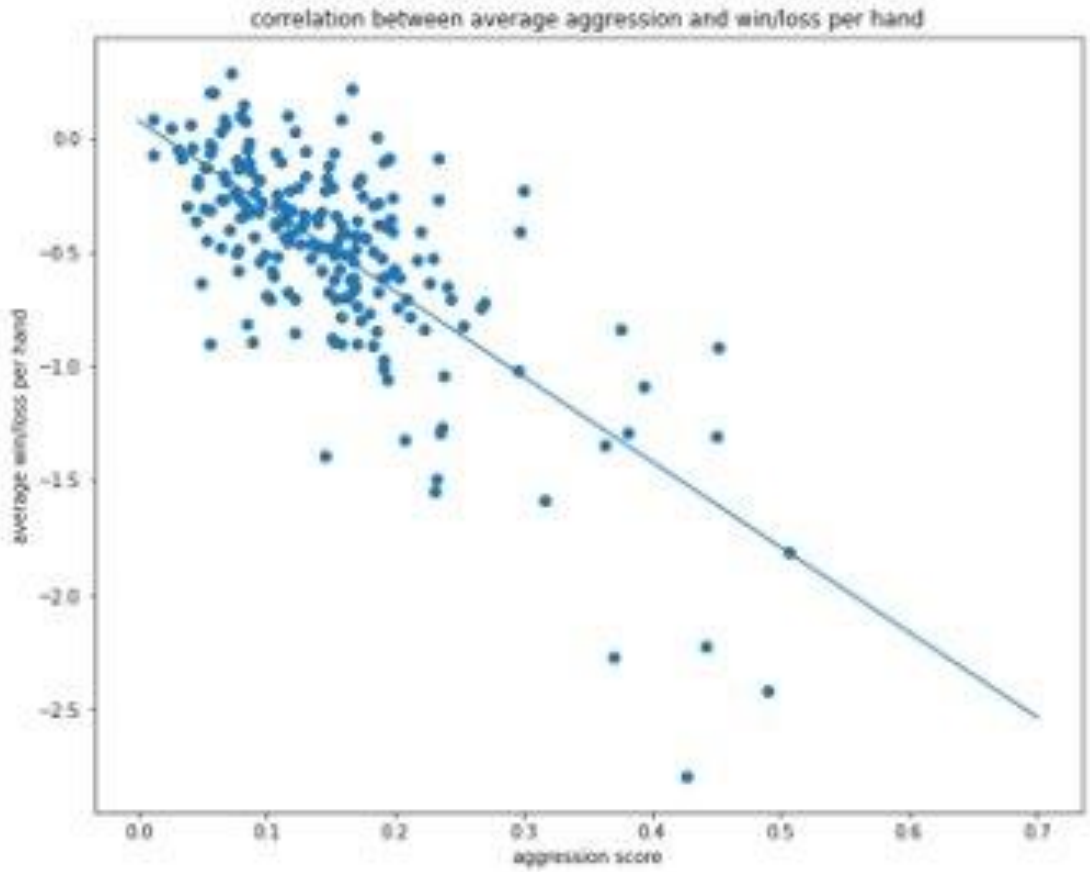

Figure 6: effect of aggression (player's willingness to put more money in pots)

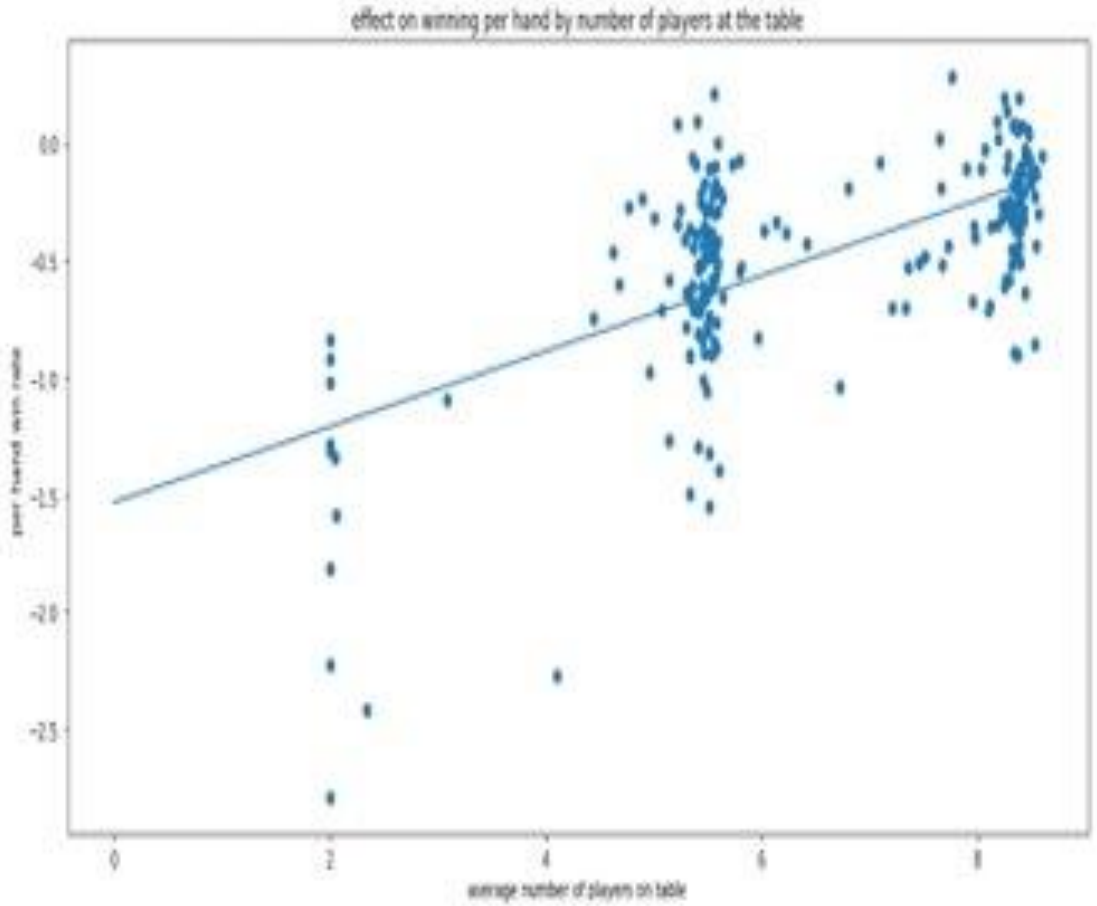

Figure 7: effect of number of players at the table

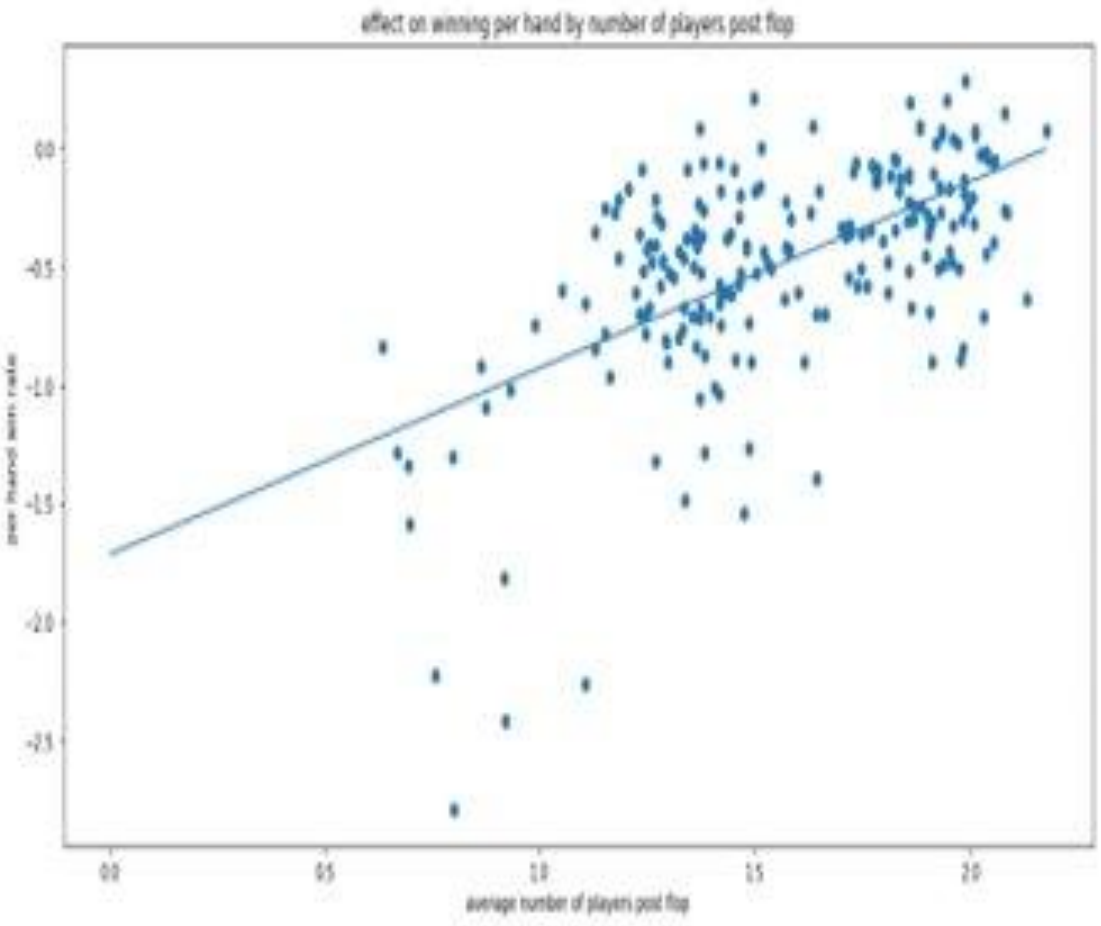

Figure 8: effect of number of players involved in hands

Figure 5 shows a positive correlation between VPIP and win rate. VPIP is a common poker analytic term, which stands for voluntarily put chips in pots, and it measures a player's willingness to be involved in pots. The player in the big blind does not count because he or she is required to put chips in pot, hence the term "voluntarily". It is commonly believed that patience is the key in poker, and a player should not be voluntarily put chips in pots without premium cards. It was somewhat surprising to see that VPIP was positively correlated with win rate. An explanation of this correlation might be that "win rate" only indicated the percentage of pots the players have won, it does not indicate the size of the pots won, so someone could have won many small pots but lost big pots and ended up a net loser.

Figure 6 shows a negative correlation between aggression and amount won or lost per hand. Aggression measures a player's tendency to perform aggressive actions such as bet and raise instead of passive actions such as check and call. The benefit of aggression in poker is to either to win a big pot with a good hand, or to bluff opponents off their hand with a bad hand. Aggression is an essential skill in poker, and being blindly aggressive can be very costly. The

negative correlation between aggression and amount won per hand indicated players were most likely unskilled in aggression, which resulted in more aggressive players lost more money.

Figure 7 showed a positive correlation between the number of players playing at the table and amount won or lost per hand. Most poker players, especially in live poker, prefer to play at a full table. A reason for this is there would be a bigger chance of encountering bad players on a fuller table. Figure 7 provided evidence for this phenomenon.

Figure 8 showed a positive correlation between number of players involved in hands and amount won or lost per hand. This correlation is often debated at low stakes poker, where playing in a hand involving more players will inevitably lower the chance of each player winning. At the same time, each player involved would be getting better pot odds, and the player who wins would win a much bigger pot in relation to investment. In this study, getting better pot odds did out weight lower chance of winning pots.

Feature selection

For feature selection, only players who played more than 1000 hands were selected, this was done to reduce the effect of random luck. This reduced the number of players from the original more than 8000, down to 329.

From 19 features of the player data frame constructed in the EDA section, 7 were chosen for model building. Amount won per hand was chosen as the label feature. Features which are highly correlated with each other were dropped. To comply with the goal of this study, selected features had to be easily understood by recreational poker players, so a set of rules could be created to improve their strategies. The selected features were the following: preflop aggression, flop aggression, turn aggression, river aggression, average pot size, number of players involved, and VPIP.

All 7 selected features were categorized. For example, aggression features were categorized by assigning a threshold for aggressive and passive. The thresholds for each feature were chosen based on the distribution of values with the feature to ensure balance. Finally, one hot encoding was done on all selected features. Figure 9 shows the head of the machine learning ready data frame.

	p_agg_aggressive	f_agg_aggressive	t_agg_aggressive	r_agg_aggressive	vpip_tight	pot_size_large	post_num_many	win_per_hand
0	1	0	0	0	1	1	0	0
1	0	1	1	1	1	1	0	1
2	0	0	0	0	1	0	1	1
3	1	1	0	0	1	1	0	0
4	1	0	0	0	1	0	0	0

Figure 9: data frame after one-hot-encoding selected features

Model selection

Python's SKLEARN module was used for majority of the modeling process. Random forest, logistic regression, and gradient boosting methods were used for modeling. All models were trained on 20/30 train test split of the available data, and 5 folds cross validation. ROC-AUC score was used for the primary evaluation metric, accuracy score was also used as a metric.

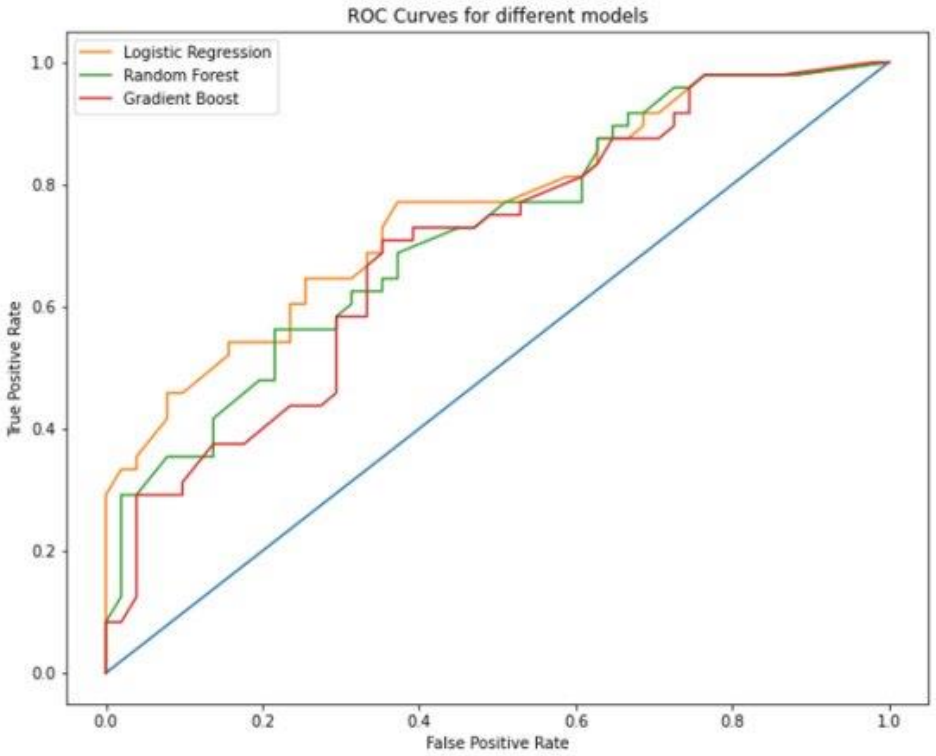

Figure 10: ROC-AUC score of the 3 selected models

Logistic regression model was selected due to slightly higher ROC-AUC score. The model was then tuned to have C of 0.1 and ridge(l2) regularize for optimal score. The tuned logistic regression model had ROC-AUC score of 0.685, and accuracy score of 0.689 on the test data. This score wasn't particularly high, mainly due some limitations such as lack of sufficient data and some potential factor that were not analyzed. This will be discussed in more detail in the conclusion section. Figure 11 shows the confusion matrix plot of the tuned model on the test data.

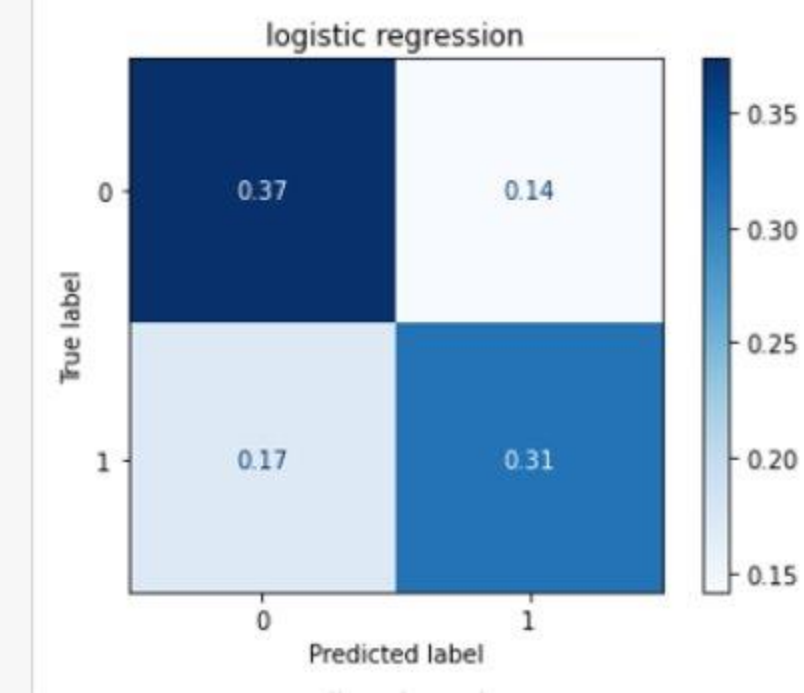

Figure 11: confusion matrix of the tuned logistic model

Finally, the goal of this study was to find what features contribute to a well performing player. Feature importance plot will help to build a set of rules for minimizing recreational played losses.

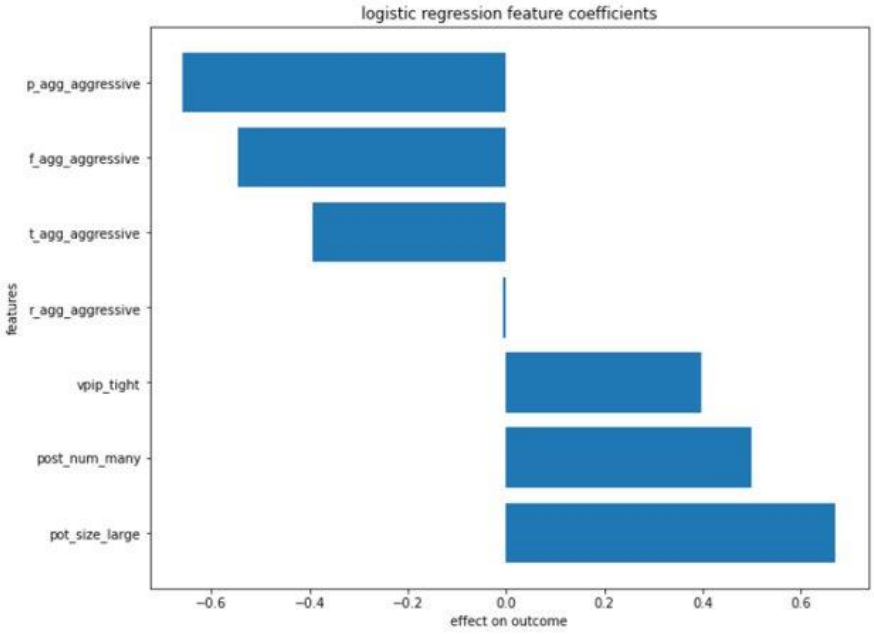

Figure 12: feature importance plot

From the feature importance plot, it was observed while playing large pots, playing in multi-way pots, and playing at a tighter VPIP positively contribute to winning per hand. Intuitively, these positive factors all make sense, such as playing in large pots and in multi-way pots, both would provide good pot odds, which allowed players to receive larger return on investments. Playing tighter also means playing higher quality cards overall, which provided better chance of making the best hand.

Also, from the feature importance plot, it appeared that being aggressive on any street negatively contributed to winning per hand. This indicated that at lower stakes, players tend to either bluff too often or betting too big to be paid off. According to this study, recreational players should be playing passively across all streets, until well-studied on correct aggressiveness.

Conclusion

Poker hand history data was gathered from an online poker site, with over 300,000 hands, over 8,000 players, and in a span of 23 days. The goal was to draw insight from these recorded poker hands, and determine what features contributed to winning players, and create a set of rules for recreational players to improve their performance at the game of poker.

Throughout the analytical process, over 20 features were created from the original text files and during the EDA process. After feature selection and feeding the selected feature to 3 models, the logistic regression model was the most accurate and consistent. The final turned logistic regression model has ROC-AUC score of 0.69 on the test data set, derived from 70/30 train/test split. 6 features were deemed important by the model, and were used to created the following rules for recreational poker players:

1. Player more large pots
2. Play more multiway pots
3. Play less than 30% of the hands dealt
4. Play passively on all streets.

As a disclaimer, the set of rules above only serve to help recreational players to lose less money, it will most likely not make a recreational player a winning player. One should only gamble what he or she can afford to lose.

There were several limitations to this analysis. First is that players' actual cards were not recorded consistently, which means it was not possible to determine the bluffing frequency of each player. Also, the poker hands were from a span of only 23 days, with most players played 1,000 hands of poker or less, so random luck may had been a factor to many players' performances. Lastly, there are some factors that would absolutely affect players' performances but impossible to analyze, such as emotion, or luck.

For future analysis on similar topics, one or more of the following improvements will be recommended:

1. Obtain more hand history per player
2. Attempt to analyze hand by hand instead of player by player
3. Attempt to analyze toughness of games, such as how many winning players are present on a table.
4. Attempt to analyze the network effect of the player pool, such as how often each player played against another player.