Springboard Capstone 2 Final Presentation

ANALYSIS OF TEXAS HOLD'EM POKER STRATEGY

The Problem

- Winning poker strategies are very complex, which involves the odds, combinatorics, observation, and self-control.
- Recreational players enjoy playing poker, but do not have time nor the dedication to develop a winning strategy.
- The goal is to find a simple strategy that players can pick up and use immediately at the poker table.

Who Cares

- Recreational poker players who intend on saving money in the game.
- Poker room owners/online poker sites who will benefit from their players stay in the game for longer periods of time.

Poker Myths and Common Beliefs

- The notion of "I have been unlucky for the past few days, so I will be lucky today". (gambler's fallacy)
- Poker is all about bluffing.
- Reading other people's body expressions is a very important skill to have.

Logical factors that affect win rate

- The overall quality of hands the player plays, where higher quality of hands should win more
- Willingness to put money in the pot, this usually indicates the bluffing frequency, which is highly dependent on skill.
- Number of players in a hand, where more players will result in a lower chance of winning.

Data Information

- Records of all the poker hands played from 07/01/2009 until 07/23/2009 on an online poker site.
- Low stake poker hands ($0.5/$1 blinds) are chosen because it's the most likely stake to find recreational players.
- 324011 poker hands recorded during that period.
- Over 8000 players were involved.
- Poker hands were recorded across 325 text files.

Data Wrangling part 1

- The original data were recorded in descriptive format in .txt files. For example, a poker hand was recorded like this:

```
Stage #3017237436: Holdem  No Limit $1 - 2009-07-01 00:00:09 (ET)
Table: INDIANA ST (Real Money) Seat #5 is the dealer
Seat 5 - vETYfpoA+FhBercnDPJrRw ($197 in chips)
Seat 6 - DeZAZcPNNQ5w+Wb+5ujZdA ($200.30 in chips)
Seat 2 - AiiJXMM0CfYl69+Nq3jyfA ($78.50 in chips)
Seat 3 - id+sbECX+YdI8qhMhpje+g ($81.60 in chips)
DeZAZcPNNQ5w+Wb+5ujZdA - Posts small blind $0.50
id+sbECX+YdI8qhMhpje+g - Posts big blind $1
*** POCKET CARDS ***
vETYfpoA+FhBercnDPJrRw - Folds
DeZAZcPNNQ5w+Wb+5ujZdA - Raises $2.50 to $3
id+sbECX+YdI8qhMhpje+g - Folds
DeZAZcPNNQ5w+Wb+5ujZdA - returned ($2) : not called
*** SHOW DOWN ***
DeZAZcPNNQ5w+Wb+5ujZdA - Does not show
DeZAZcPNNQ5w+Wb+5ujZdA Collects $2 from main pot
*** SUMMARY ***
Total Pot($2)
Seat 3: id+sbECX+YdI8qhMhpje+g (big blind) Folded on the POCKET CARDS
Seat 5: vETYfpoA+FhBercnDPJrRw (dealer) Folded on the POCKET CARDS
Seat 6: DeZAZcPNNQ5w+Wb+5ujZdA (small blind) collected Total ($2)
```

- Some of the information to be extracted from the text:
  - Hand ID
  - Player ID
  - Stack size
  - Position
  - Preflop, flop, turn, river actions and amounts

Data Wrangling Part 2

- The primary tool used to extract information from the text files is python's REGEX module.
- Temperorily stored information in lists for the next step of processing.
- Example code to extract each player's amount of money invested preflop from the text file is:

```
pre_amount = []
for n, i in enumerate(pre_flop):
    x = active_players[n]
    y = re.findall(r'(.{22}) \- (?:Raises|Checks|Calls|Bets) \$(\S+)', i)
    amounts = []
    for a in x:
        amount = 0
        for b in y:
            if a == b[0]:
                amount += float(b[1])
        amounts.append(amount)
    pre_amount.append(amounts)
len(pre_amount)
```

Data Wrangling Part 3

- Python's PANDAS dataframe was used to record the extracted data.
- Combined all the lists constructed in the previous step into a dataframe.
- Each row indicated each player in each hand
- Resulting dataframe was 1767588 rows by 16 columns

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

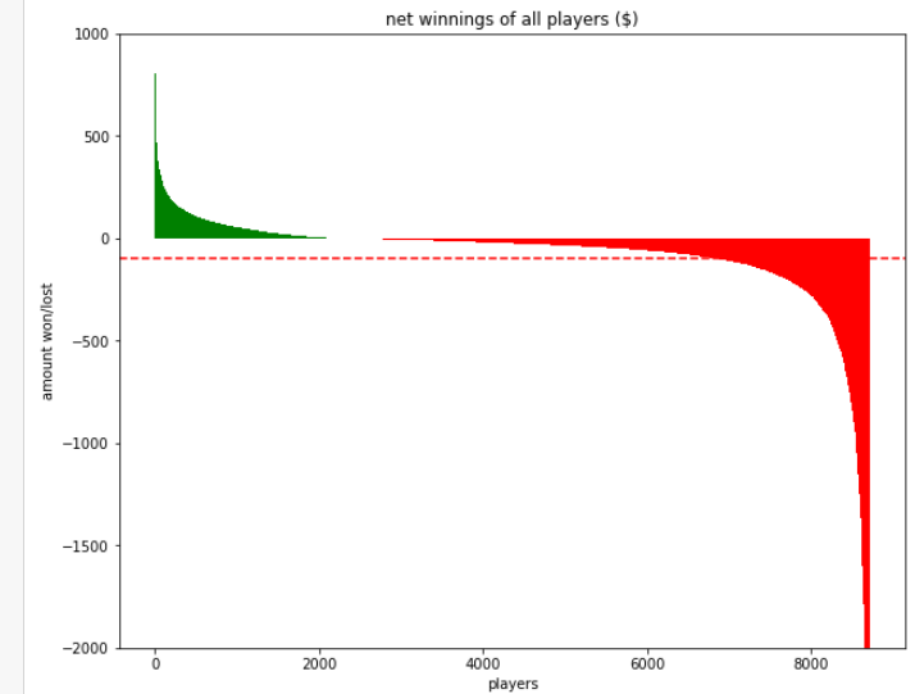
EDA Part 1

- The analysis focused on individual players instead of individual hands.
- Dataframe from the previous slide were grouped by players.
- Features were created per player, such as hands played, net amount won, VPIP, average pot size etc.

	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
Ir5ondonH45KZUPlsjjOZg	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

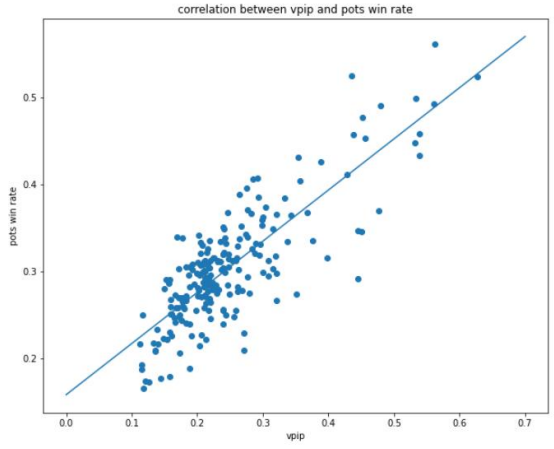
EDA part 2

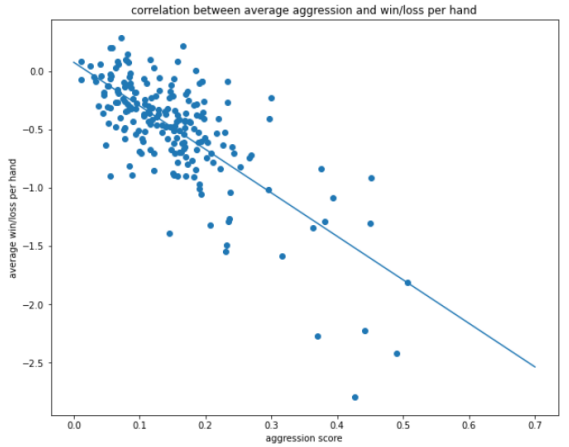
- How was the overall win/loss of all players.

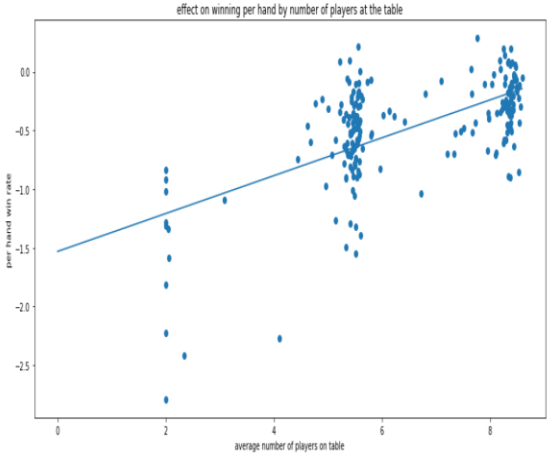


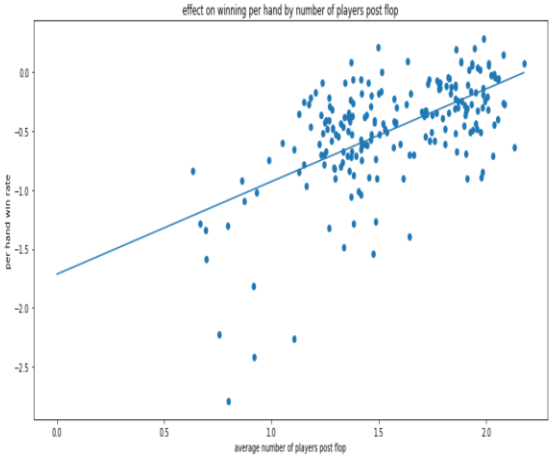
- Following were observed from the above graph:
  - Over 8000 players were involved.
  - Many more losing players than winning players.
  - Losing players lost much more than winning players won.
  - Average was around –80 instead of 0, indicated the effect of rake.

EDA part 3









Some example plots of different features vs win rate.

- VPIP was possitively correlated
- Average aggression was negatively correlated
- Number of players at the table is positively correlated
- Number of players post flop was positively correlated

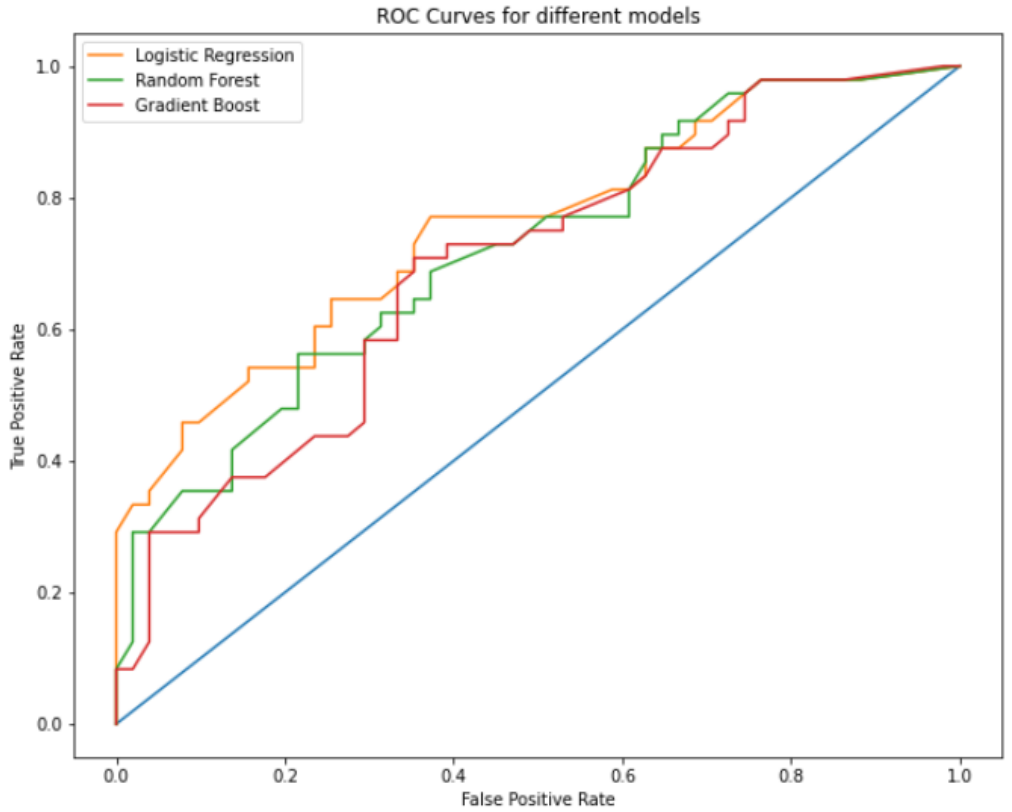
Feature Engineering

- From more than 8000 players involved, only players who played more than 1000 hands are selected, resulting in 329 players.
- From 19 features, 7 were selected, and winning per hand was selected as the label.
- All variables were represented with binary categories, and then encoded into 1s and 0s

	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

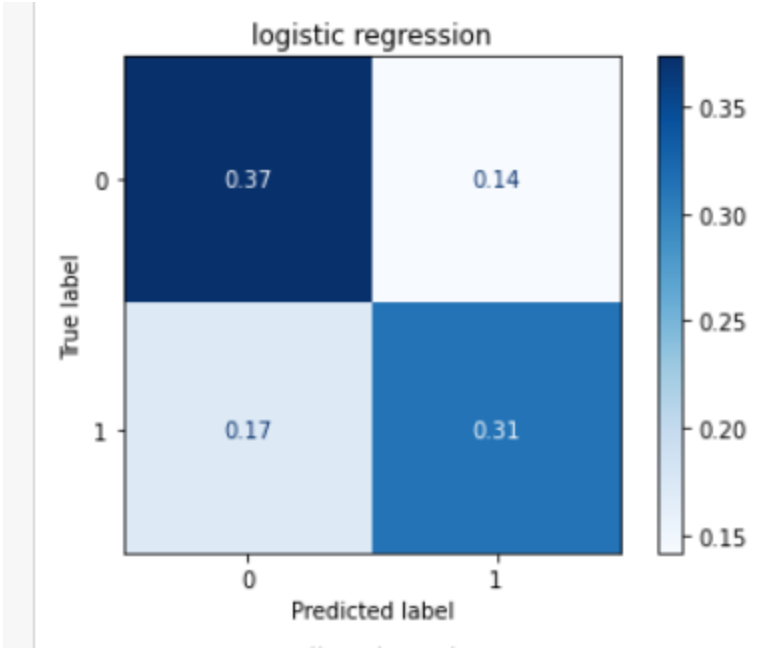
Modeling part 1

- Python's sklearn module was used for the majority of the modeling process.
- Used 70/30 train test split, 5 folds cross validation.
- Tested 3 models: random forest, logistic regression, and gradient boosting.



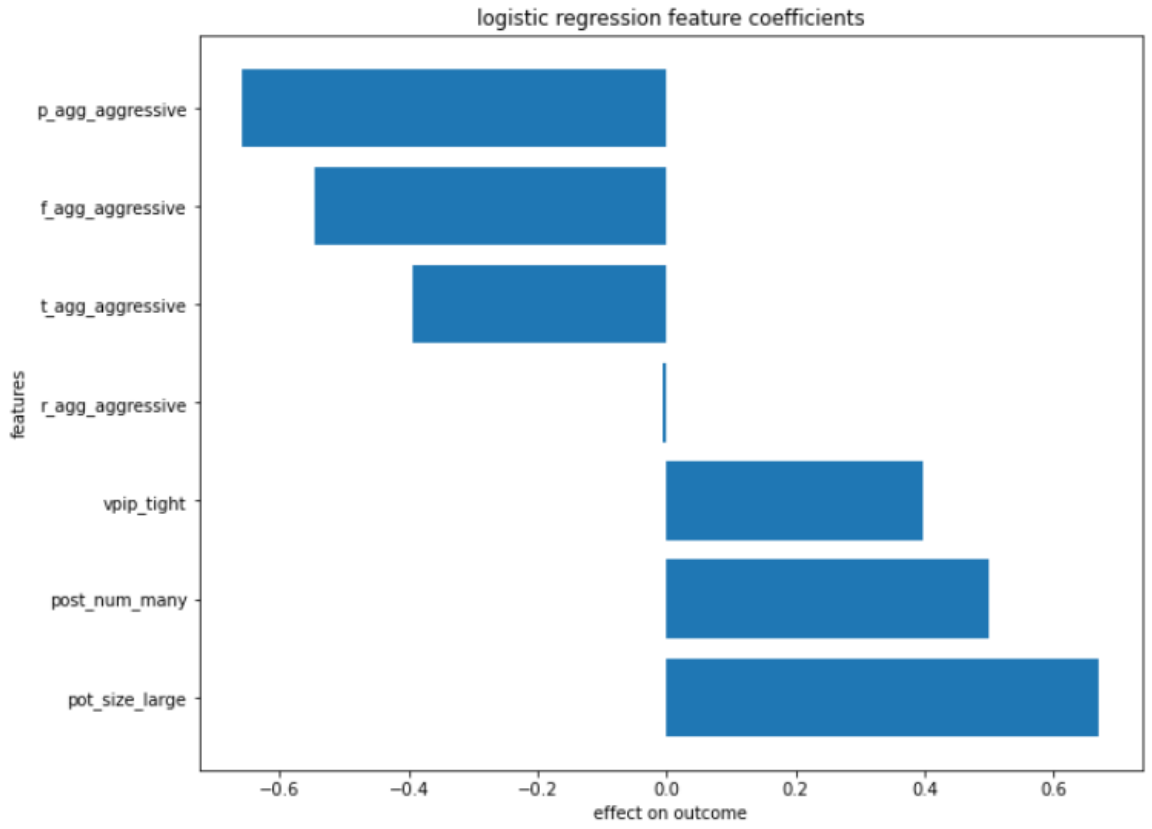
Modeling part 2

- Chose the logistic regression model due to slightly higher ROC-AUC score
- Tuned logistic regression model resulted in C of 0.1 and ridge(l2) regularizer
- Model had ROC-AUC score of 0.685, and accuracy score of 0.689 on the test data
- Confusion matrix plot of the test data:



Modeling Part 3

- How does each feature affect the win rate?
- Feature importance plot will help to build a set of rules for minimizing recreational players' losses.



Limitations and Disclaimers

- Limitations:
  - Players' actually cards were not recorded, which means it was not possible to test the impact of hand quality.
  - The poker hands were from a span of only 23 days, more data would have resulted more accurate modeling.
  - A few factors that were hard to analyze, such as luck and players' emotions.

- Disclaimer: The set of rules found by the model only serves to help recreational players to lose less money, it will most likely not make a recreational player a winning player.

Ideas to improve future modeling

- Obtain more hand history per player.
- Attempt to analyze hand by hand instead of player by player.
- Attempt to analyze table envoriment, such as how many winning players are present on a table
- Attempt to analyze the network effect of the player pool, such as how often player A played on the same table as player B.

Conclusion

- Over 20 features are created from the original text files and during the EDA process, only 6 features were deemed important by the model in the end.
- Out of the 3 tested supervised learning models, the logistic regression model was the most accurate and consistent.
- The final tuned model has ROC-AUC score of 0.69 on the test data set, derived from a 70/30 train/test split.
- The set of rules recreational players should follow are:
  - Play more large pots
  - Play more multiway pots
  - Play less than 30% of the hands dealt
  - Play passively on all streets