Predictive Modeling for Rounds of Blackjack

Mohil Patel
Penn State University
State College, PA
I am a senior in Computational Data Science, and I hold a strong interest in sports data analytics and business intelligence. I believe my expertise is in data visualization and data modeling.

Sagan Kakkar
Penn State University
State College, PA
I am a fourth-year Data Science and Statistics major who has an interest in real-world data and statistical analysis via advanced machine learning methods. My expertise lay in the field of statistical analysis and model creation.

Abstract— **This project aims to predict the next move a player should make when playing the card game Blackjack. A variety of data science techniques were used to collect data, analyze data, and make predictions. Traditional strategies for playing blackjack often rely on basic rules and heuristics, neglecting the potential advantages offered by advanced predictive modeling. Leveraging a dataset of simulated blackjack games and outcomes, we employ machine learning algorithms to develop a predictive model capable of anticipating player and dealer decisions. Insights gained from the model's decision-making process contribute to a deeper understanding of optimal blackjack strategies, shedding light on nuanced patterns within the game. Furthermore, the paper discusses potential implications for both casino gaming and player strategy, emphasizing the role of data science in enhancing decision-making processes in games of chance. As the field of predictive analytics continues to evolve, this research represents a significant step toward leveraging machine learning to optimize decision-making in strategic card games like blackjack.**

Introduction

Blackjack is a widely enjoyed card game known for its blend of strategy and chance. The challenge lies in making informed decisions based on the cards dealt to the player and the dealer. The goal of the game is for the player to get as close to 21 as possible without going over and at the same time get a better hand than the dealer. If a player can make correct decisions based on the cards, he/she can be the subject of massive amounts of winnings. Vice-versa, if the player makes the wrong decision, they can lose a substantial amount as well. Leveraging the power of machine learning techniques, we aim to create a predictive model that can forecast the ideal move to make based on card information that a player would typically have. Collecting a robust and diverse dataset is pivotal for the model's effectiveness. We will simulate a substantial amount of data due to the lack of detailed record keeping for blackjack games. The data will encompass details like player and dealer cards as well as hand outcomes. This data will serve as the foundation for training and assessing the predictive model.

Related Work

We have compiled a comprehensive collection of prior work related to this project and will discuss the main ones and how they relate to our project. The source that we found to be the most similar to our project is an article by Coder's Haven. This source outlines the use of XGBoost in Python in order to create a predictive model and compare it to a model representing basic strategy and a model representing randomly selecting cards. The work done in this project gives us a solid outline of the general process and an idea of what the results may look like.

The next source we found to be useful is a paper that is a part of the International Research Journal of Engineering and Technology. We found this to be very interesting because it covers a different modeling approach to our goal of creating a predictive blackjack model. In this paper, reinforcement learning and Artificial Neural Network (ANN) are used in order to create a model that actually simulates human behavior. According to Khalkhar, “This game-playing bot's activities will never be limited to standard steps taken by players, but will instead create its own methods and new variations, making it impossible for a player to simply memorise its moves and beat it” (2022). This piqued our interest and gave us an outline for a different approach to our goal and new ideas of how to approach the situation.
In the article, “Winning Blackjack using Machine Learning” (Sommerville 2019), the basic strategy of blackjack is covered along with the basic outline of how algorithms can try and mimic the game. We used this article mainly as an introduction document to get ourselves thinking of different unique ways to approach the topic. This article also gave us quality information concerning basic strategy that we will use in order to build our model that represents basic strategy.

The next paper covers the probability associated with the game of blackjack. As stated in his paper, “the purpose here is to calculate results for a typical hand where sampling without replacement is employed. It is seen that significant error can result when long runs are required to complete the hand” (Cooke 2010). The mathematical formulas and methods shown in this paper are important for our work because we can use the basic mathematical principles behind the game in order to try and make our model use these principles in the most efficient way for predictions.

In a paper from Stanford University, a similar project to what our goal was detailed. The authors used various methods to try and optimize blackjack strategy and three were deemed to have performed better than basic strategy. As stated, "Value

Iteration, Sarsa, and Q-Learning, resulted in policies with win rates far superior than a random policy when considering the agent's hand" (Geiser 2020). We found this important because it gives us insight on different successful ways to approach the same question: how to optimize blackjack play via models? We also found this particular paper very appealing because the paper includes the code that was used in their project. We plan on using the fundamental ideas of the provided code in our project in order to test if these methods will prove effective for our project.
Lastly, in a paper from Rice University the authors detail Markov Analysis on the game of Blackjack and the mathematical advantages or disadvantages that could occur through the course of the game. As stated in the paper, "Exploiting this structure and elementary results from the theory of Markov chains, we present a novel framework for analyzing the expected advantage of a card-counting system entirely without simulation" (Wakin & Rozell). This shows yet another unique way to approach the question mentioned numerous times before. We believe that this is important because it allows us to approach the topic from a more theoretical stance rather than just a technical one.

In short, the above papers give us a solid foundation of information that we will use in our project in order to create the model that predicts the correct move to make based on table card information. The papers vary in their information from theoretical to technical. Having a quality mix of both will allow us to approach this problem with a knowledge base ideal for success.

Methods

Dataset

We have utilized ChatGPT to simulate a dataset encompassing over 1000 rounds of blackjack. We decided to simulate the data instead of using an online source because there are few, if any, databases that record data round by round that we desired for our project. This is the case due to privacy reasons and security reasons of casinos. Therefore, we decided simulation using ChatGPT would be the most efficient method and would allow us to customize our data set while keeping it random at the same time. One thing to note as we discuss our data is that a player win is when the sum of the player's cards is closer to 21 than the dealer's are or the dealer goes over 21. A dealer's win is when the dealer's cards are closer to 21 than the player's cards are, or the player goes over 21 (busts). We have tried multiple different formats of a potential dataset and decided on a format like the sample shown below.

Round	Player's Card 1	Player's Card 2	Dealer's Card 1	Dealer's Card 2	Player's Action	Dealer's Action	Outcome
1	7	5	9	6	Stand	Hit	0
2	10	8	J	5	Stand	Stand	1
3	A	7	10	2	Stand	Hit	0
4	6	9	A	3	Stand	Hit	1
5	3	4	7	A	Split	Stand	0

As seen above, this dataset gives the model a whole column of extra information that will help with the prediction. This data set also gives a wider variety of moves the player can make, such as splitting and doubling down. These moves are not as common in the game of Blackjack yet are important to know and be able to use when necessary. Because of this, we will try to incorporate these moves in our prediction. In total, there are 1094 rounds of simulated data that we will use for our project's next phase.

Data Preprocessing

We used libraries such as pandas for data manipulation, sklearn for model implementation and evaluation, and xgboost for utilizing the XGBoost classifier. We ensured the dataset was clean, null values were handled and categorical variables were encoded to numerical ones to make them machine-readable. We mapped the card values to their respective numerical values and encoded the players' and dealer's actions using Label Encoding. The "Player's Action 1" and "Dealer's Action 1" columns were transformed into numerical formats to facilitate the training of machine learning models. We mapped the card values to their respective numerical values and encoded the players' and dealer's actions using Label Encoding. The "Player's Action 1" and "Dealer's Action 1" columns were transformed into numerical formats to facilitate the training of machine learning models.

Feature Selection and Splitting

The dataset was divided into features (X) and the target variable (y), with 'Round' and 'Result' columns being excluded from the features. The data was then split into training and testing sets, with 80% allocated for training and 20% for testing, using a random seed of 42 to ensure reproducibility.

Model Training and Evaluation

Several classification models were employed, including XGBoost, Random Forest, Support Vector Machine, and Logistic Regression. These models were trained using the training data and evaluated on their accuracy in predicting the test data. The best performing model was selected based on the highest accuracy score.

Hyperparameter Tuning

For the XGBoost model, a RandomizedSearchCV was conducted with specified parameter grids to optimize the model's hyperparameters. The process aimed at maximizing accuracy. The best parameters were then applied to train the final XGBoost model.

Automated Machine Learning (AutoML)

An AutoML approach was also undertaken using the flaml library, focusing solely on the XGBoost estimator. The AutoML's task was set to 'classification', with a time budget of 120 seconds and a goal to optimize for accuracy. The best model was saved, and its performance was evaluated against the test dataset.

Evaluation

The performance of each model trained on the dataset is summarized below:

- XGBoost: Achieved an accuracy of 83.56%. This model showed a strong performance and was the initial choice before optimization.
- Random Forest: Resulted in a lower accuracy of 71.69%, indicating that this model was less effective at capturing the patterns in the dataset.
- Support Vector Machine: Had an accuracy of 66.67%, suggesting that this model might not be well-suited for this particular dataset.
- Logistic Regression: Reported an accuracy of 65.75%, which was the lowest among the models tested and implies limited predictive capability for this dataset.

Optimized Model Performance

An optimized XGBoost model, fine-tuned through RandomizedSearchCV, demonstrated an accuracy of 80.37%. Although this is a slight decrease from the initial XGBoost model, it indicates the nuances of hyperparameter tuning and its impact on model performance.

AutoML Results

The AutoML process using the flaml library focused on the XGBoost estimator and yielded the following outcomes:

- Best Model: XGBoost was identified as the best machine learning learner.
- Hyperparameter Configuration: The best hyperparameter configuration included settings like n_estimators: 21, max_leaves: 100, learning_rate: 0.2417810635584405, etc.
- Validation Accuracy: The best accuracy on the validation data was approximately 82.45%, which is a notable result given the constraints of time and computational resources.

Final Model Evaluation

The final model, as determined by AutoML, achieved an accuracy of approximately 83.11% on the test dataset. This accuracy is comparable to the initially trained XGBoost model and showcases the effectiveness of AutoML in identifying competitive models under a constrained setting. The experimentation with various models and techniques demonstrated that the XGBoost algorithm, both in its original and optimized forms, was particularly effective for this dataset. The AutoML approach was beneficial in fine-tuning and selecting an appropriate model that balanced performance and computational efficiency.

Discussion

After experimenting with various algorithms, it is clearly seen that the XGBoost worked well for our intended predictions. At the beginning of the project, we were not sure how well ChatGPT would be at simulating the data, but it turned out to produce quality results that were usable by our model. One thing that did not work as we wanted was the UI that would allow users to input their cards and the dealer card and then receive an output of what move to make. While we have the basic structure of this interface, we were not able to get it working to our standards yet.

Ethical Considerations:

In developing and applying predictive models for games like Blackjack, several ethical considerations were taken into account:

1. Responsible Gaming: The model's potential to enhance player performance raises concerns about fair play and responsible gambling. It's important to ensure that the use of such predictive tools does not lead to addictive behaviors or provide an unfair advantage over unaided players or the house. The promotion of responsible gambling practices should accompany any application of this technology.
2. Data Privacy and Security: Since the model requires player and game data, it's crucial to handle this information responsibly. Ensuring data privacy and security, especially in an industry where financial transactions are involved, is of paramount importance. This includes complying with data protection regulations and ensuring that user data is not misused or accessed without consent.

3. Transparency and Accountability: The use of AI and machine learning in gambling should be transparent. Players should be informed about how their data is being used and how the model makes predictions. Moreover, developers and operators should be accountable for the model's impact, especially in cases of error or bias.
4. Equity and Access: The availability of such predictive tools should not create a divide between players who have access to them and those who do not. It's important to consider the broader implications of technology that could potentially skew the playing field in favor of those with access to more advanced tools.

Business Implications:

The integration of predictive modeling in Blackjack presents several business implications:

- Casino Management and Operations: Casinos could use the model to refine their game management strategies, potentially leading to increased profitability. However, they also need to balance this with the maintenance of fair play and customer trust.
- Player Strategy Tools: There's a market for player-oriented applications that use predictive models to advise on game strategies. This could be particularly appealing to new players or those looking to improve their skills.
- Regulatory Impact: As AI becomes more prevalent in gaming, regulatory bodies may need to update or create new guidelines and rules to govern its use. This could affect how casinos and gaming platforms operate and market their services.
- Competitive Advantage: For casinos and online gaming platforms, leveraging AI for game analysis and customer engagement can provide a competitive edge. However, this also raises questions about technology-driven disparities between different operators.

While the application of predictive models in Blackjack offers opportunities for innovation and profit, it also necessitates careful consideration of ethical, legal, and market dynamics to ensure sustainable and responsible business practices.

Conclusion

Overall, we term this project a success and a step in the right direction to a bigger goal. Our intention was to create a process that could be reasonably accurate in its predictions for the next move in a game of Blackjack, given the player's cards and the dealer's card. While we would have liked to generate a higher accuracy score, the accuracy of 83.56% is satisfactory given that this is our first iteration of the model. This model is useful for those in the gaming industry, for it gives these individuals a tool to get an idea of ideal player strategy and moves given a set of cards. This model could also be useful for those wanting to learn Blackjack and what moves to make before they play it.

Moving forward, we would first like to make some alterations to our current code to see if we can improve the accuracy of the current models. These changes may take the form of hyper-parameter tuning or other similar tweaks that could have a significant effect on the accuracy of the models. Also, we would like to continue working on our User Interface for predicting the next move if given the player's cards and dealer's card. We have done a fair amount of coding and created a baseline user interface. However, we have run into some issues and the output is not what we expect or desire. Going forward, we would like to get this working to a point where the user can input their cards and a dealer card and then be given the move that they should make next. For example, the user could input A, 9 for their cards and 6 for a dealer card and then hit a button which says predict which will ideally display the correct move the player should make – in this case it should display "Stay".

Contribution

Sagan Kakkar: I was responsible for doing preliminary research and finding related works, simulating the dataset through ChatGPT, conducting data exploration through visualizations and numerical analysis, and drafting reports to display information. The code written for the data exploration can be found in our GitHub and is labeled VisualizationCode.Rmd.

Mohil Patel: I was primarily responsible for the coding part of the project which included model development and integration, data preprocessing and analysis, hyperparameter tuning and optimization, code documentation(Github) and management. I aimed to ensure that the technical foundation of our project was robust, efficient, and capable of achieving our goal of developing an accurate predictive model for Blackjack.

Citations

(2020).Hackathon: Training a Blackjack AI. www.codershaven.com/posts/hackathon-training-a-blackjack-ai/.

ChatGPT, GPT-4, OpenAI, 30 Jun. 2023, chat.openai.com/chat.

Cooke, H. C. (2010). Probability models for blackjack poker. Computers and Mathematics with Applications. 59(1). doi:108-114. doi:10.1016/j.camwa.2009.08.062.

Geiser, J & Hassler, T. (2020). Beating Blackjack – A Reinforcement Learning Approach. https://web.stanford.edu/class/aa228/reports/2020/final117.pdf.

Khalkar, R, et al. (2022).Gaming Bot Using Reinforcement Learning. International Research Journal of Engineering and Technology (IRJET). 9(7). www.researchgate.net/profile/Rohini-Khalkar/publication/371576740_GAMING_BOT_USING_REINFORCEMENT_LEARNING/links/648ae5fc7fcc811dcdceee60/GAMING-BOT-USING-REINFORCEMENT-LEARNING.pdf?origin=publication_detail.

Sommerville, G. (2019). Winning Blackjack using Machine Learning. https://towardsdatascience.com/winning-blackjack-using-machine-learning-681d924f197c.

Wakin, B. M & Rozell, J. C. A Markov Chain Analysis of Blackjack Strategy. https://inside.mines.edu/~mwakin/papers/mcbj.pdf.