MITIGATING RISKS FOR THE VIRGINIA LOTTERY

Farhan Khan
crg3ts@virginia.edu

Zohaib Khalid
cxk3hk@virginia.edu

Ben Doniger
kpg9cj@virginia.edu

December 13, 2024

1 Abstract

This project investigates lottery scratcher games across multiple U.S. states to identify those that offer players better chances of winning, aiming to empower more informed and responsible participation. Utilizing the "Predicting the Lottery" dataset from Kaggle—which encompasses detailed information such as ticket prices, prize amounts, odds, and the percentage of remaining winning tickets from California, New Mexico, Missouri, Oklahoma, and Virginia—we approach the analysis as a regression problem. The odds of winning serve as the target variable for our predictive models.

We employ various machine learning regression algorithms, including linear regression, gradient descent, and random forest regression, to predict the odds of winning prizes. The performance of these models is evaluated using the root mean square error (RMSE) metric to determine their predictive accuracy. Our approach distinguishes itself from existing platforms like ScratchSmarter by leveraging advanced machine learning techniques rather than relying solely on statistical models and real-time data tracking.

The ultimate goal is to pinpoint games that offer the best return on investment for players and to provide actionable recommendations. By enhancing transparency and offering data-driven insights, this research aims not only to improve individual decision-making but also to contribute to more equitable game design. Additionally, it has the potential to address broader issues such as gambling addiction by promoting responsible gaming choices informed by robust data analysis.

2 Introduction

2.1 Motivation

Our project focuses on analyzing data from a wide range of lottery scratcher games to determine which games provide players with a better chance of winning. The motivation behind this study is to offer players a more informed opportunity when participating in these games. By identifying these patterns, we hope to assist in addressing broader issues as well, such as gambling addiction, by equipping players with knowledge that may help them make more responsible choices when playing these types of games. We are considering the setting of lottery scratcher games in the United States, using data from California, New Mexico, Missouri, Oklahoma, and Virginia, but focusing mainly on Virginia. In this context, players are faced with numerous game options that vary in odds and prize structures, often without transparent information to guide their choices. By analyzing this setting, we aim to uncover insights within the existing lottery systems of these states to help players navigate the complexities of these games more effectively.

2.2 Dataset

We will use a publicly available dataset on U.S. state lottery scratcher games from Kaggle, titled "Predicting the Lottery", which includes data from California, New Mexico, Missouri, Oklahoma, and Virginia. The dataset provides comprehensive information on lottery games, such as ticket prices, prize amounts, odds, and the percentage of winning tickets still available. The dataset can be accessed here.

2.3 Related Work

An example of a related work to our goals is ScratchSmarter, which is an advanced data analytics platform designed specifically for analyzing scratch-off lottery tickets. It provides detailed insights into lottery games across various states, offering a comprehensive breakdown of prize odds, ticket availability, and game performance. The platform leverages statistical models to present a clearer picture of a game's overall odds and the likelihood of winning specific prizes. Similar to our goals, ScratchSmarter's goals are related to providing transparency of lottery insights to players, striving for fairness in the games. From the information available, ScratchSmarter focuses primarily on detailed data analytics rather than explicitly using machine learning. The platform provides real-time odds updates and advanced statistical analyses of lottery scratch-off tickets, but it does not specifically mention the use of machine learning algorithms in its process. Its approach relies heavily on statistical models and real-time data tracking to help users understand prize probabilities, remaining tickets, and game performance. This is a key difference between ScratchSmarter and our project, as we use advanced machine learning models to make our analysis, unlike ScratchSmarter.

3 Methodology

3.1 Intended Experiments

We will approach this as a regression problem, where the odds of winning will be the labels to the data points. Our experiments will include training various machine learning regression algorithms to determine which can most accurately predict odds of winning a prize. The algorithms used are likely to include, but are not limited to, linear regression, gradient descent, and random forest regression.

We will evaluate our machine learning algorithms by comparing the root mean square error (RMSE) to find the closest odds prediction. Our ultimate goal is to identify which games offer the best return on investment for players and provide recommendations for players so that they can have a more fair experience with the lottery.

3.2 Research Design

Overview Our approach involves using supervised regression models to predict the odds of winning in lottery scratcher games, with a focus on regression techniques such as linear regression and random forest regression. We chose these models because they provide a balance of interpretability and predictive power. Linear regression offers a straightforward approach for identifying relationships between features and odds, while random forest regression introduces robustness and non-linearity, which may capture complex patterns in lottery odds better than linear methods alone. Additionally, gradient descent optimization will be employed to fine-tune model parameters, enabling us to incrementally improve the models' predictive accuracy.

Data Visualization To get a better understanding of the data, we made a correlation matrix to determine which features were most positively and negatively correlated to the "Probability of Winning Any Prize" label (Figure 1). The most positively correlated features were "Starting Probability of Winning Any Prize" (cor = 1.0), "Starting Probability of Winning Profit Prize" (cor = 0.711), and "Probability of Winning Profit Prize" (cor = 0.708). The most negatively correlated features were "overallodds" (cor = -0.894), "Current Odds of Any Prize" (cor = -0.894), and "Starting Odds of Profit Prize" (cor = -0.697).

We also created a scatter matrix to visualize the some of the features in the data (Figure 2). Specifically, we plotted "Probability of Winning Any Prize", "topprize", "price", "Rank by Least Expected Losses", "overallodds", and "Rank by Best Probability of Winning Any Prize."

Data Preparation To prepare the data, we preprocessed each lottery game's data by standardizing and handling any missing values, ensuring consistent feature scaling across models. We prioritized Virginia lotteries initially to capture specific state-level nuances. After obtaining satisfactory results, we will broaden our analysis to include lotteries from other states, allowing us to refine the model to accommodate diverse game structures and odds distributions.

We began by merging all relevant Virginia lottery data into a single data frame, then dropping duplicate columns. Just for preliminary purposes, we just dropped rows that contained NaN values. This comprehensive dataset provided a solid foundation for our analysis. We isolated the probability of winning any prize as our target variable y, which we aimed to predict using various features from the dataset.

Training a Basic Linear Regression model We trained a Linear Regression model on the cleaned data. Initially, the model exhibited an exceptionally low Root Mean Square Error (RMSE) of 2.6145581484178088e-08, suggesting

highly accurate predictions. However, this result prompted us to conduct error analysis, as such low RMSE might indicate overfitting. Upon closer examination, we found that certain features in the data were highly correlated with the target variable, effectively leaking information and enabling the model to predict too accurately. We confirmed that this high accuracy is due to overfitting, and adjusting the model should improve its generalizability.

Given our preliminary results, which yielded a very low RMSE with minimal data preprocessing, we refined and extended our model to improve robustness and generalization.

Implementing a Data Processing Pipeline To streamline the data cleaning process and ensure consistency, we developed a data processing pipeline. This pipeline handles missing values using a simple imputer with the median strategy. Additionally, it uses the Standard Scaler to scale all the numerical data. We discovered that the categorical data was not useful information and contained missing values. Since our data was contained in two separate tables, we removed duplicated columns that appeared due to the merge.

Evaluating leaking information Our initial results yielded a very low RMSE, leading us to believe our model was overfitting the training data. We determined that this was due to features that contained data too similar to our label: "Probability of Winning Any Prize." These features included those such as "Starting Probability of Winning Any Prize" and "Change in Probability of Profit Prize."

Training Complex Regression Models Although our linear regression model has room for significant improvement, we decided to experiment with more complex regression models like random forest regression and gradient-boosted trees. We then compared the cross-validated RMSE on the validation set to determine which model performed the best.

Hyper parameter Tuning and Model Selection After deciding to proceed with the gradient-boosted tree model due to high performance, we fine-tuned the hyper parameters. We used Randomized Search Cross Validation with 10 iterations and 5 folds to produce fast results that optimized performance.

3.3 Member Contribution

We split up the work and contributed equal amounts. Ben worked on data visualization, preparation, and creating the pipeline. Farhan worked on training the models to the test set and getting initial results. Zohaib worked on fine tuning and the final results. All members worked together on the Video Presentation and each wrote their corresponding sections in the report.

4 Results and Discussion

The XGBRegressor model was fine-tuned using RandomizedSearchCV with 5-fold cross-validation. The hyperparameter combinations were tested for 10 iterations, evaluating model performance based on negative mean squared error (MSE). The best hyperparameters were:

- n_estimators=800 (number of boosting rounds)
- max_depth=3 (tree depth controlling complexity)
- learning_rate=0.05 (step size for optimization).

Higher max_depth values (6 or 9) and extreme learning_rates (0.1 or 0.01) often resulted in higher RMSE, suggesting overfitting or underfitting, respectively.

Our XGBRegressor model resulted in a final RMSE of 0.0156 on the holdout test set, demonstrating the model's effectiveness in reducing error and improving predictive accuracy. In terms of feature importance, the features with the most dominant influence on predicting the odds of winning were overall odds (68.8%) and ticket price (30.5%). This indicates that these variables play a critical role in determining game outcomes and return on investment.

5 Future Work

Once we have a refined model for Virginia lottery games, we will apply it to additional states in our dataset, such as California, Missouri, New Mexico, and Oklahoma. This step will test the model's adaptability to different state-specific games and determine if unique patterns emerge across states. This will also help remove bias to just VA lottery rules and improve generalization as a whole.

6 Conclusion

Our analysis demonstrates that using advanced machine learning models, like XGBRegressor, can help to accurately predict lottery odds, achieving an RMSE of 0.0156 on the Virginia lottery data. This validates our hypothesis that machine learning offers a more effective approach than traditional methods for understanding lottery games. The results also emphasize the importance of features like overall odds and ticket prices in shaping outcomes, providing meaningful insights for players aiming to make informed decisions.

This project has important implications for the well-being of Virginia and its residents. By promoting transparency and empowering lottery players with data-driven insights, we aim to encourage more responsible participation in lottery games to help curb addictions. These findings can also guide state lottery organizations toward creating games with fairer odds, ultimately building trust with participants and fostering a more equitable and healthier gaming environment.

That said, there are some limitations to consider. Since this study primarily focused on Virginia, the results may not yet fully apply to states with different lottery structures. Additionally, the initial data leakage that we encountered highlights the need for more thorough data pre-processing to ensure robust and reliable results.

Looking ahead, expanding our analysis to include other states will allow us to test the adaptability of the model and uncover broader patterns in lottery games. Incorporating real-time data tracking is another exciting direction that could make our predictions even more relevant and actionable. By continuing to refine and expand this work, we hope to contribute to fairer lottery systems and help players make smarter, more informed choices.

7 Figures

Out[9]:

Probability of Winning Any Prize	1.000000
Starting Probability of Winning Any Prize	1.000000
Starting Probability of Winning Profit Prize	0.711708
Probability of Winning Profit Prize	0.708187
topprize	0.591259
StdDev of All Prizes	0.584905
price	0.539772
Expected Value of Profit Prize (as % of cost)	0.376666
Expected Value of Any Prize (as % of cost)	0.376338
StdDev of Profit Prizes	0.334566
Rank by Best Change in Probabilities	0.270829
Rank by Least Expected Losses	0.257383
Current Odds of Top Prize	0.242498
Rank by Most Available Prizes	0.236342
Days Since Start	0.230873
Overall Rank	0.193874
Change in Expected Value of Any Prize	0.187070
Change in Expected Value of Profit Prize	0.187070
Rank Average	0.177208
Percent of Profit Prizes Remaining	0.155315
Percent of Prizes Remaining	0.153026
Percent Tix Remaining	0.150513
Change in Current Odds of Any Prize	0.077182
Total remaining	0.001274
Change in Probability of Any Prize	-0.000823
Ratio of Decline in Prizes to Decline in Losing Ticket	-0.011137
Change in Current Odds of Top Prize	-0.036764
Winning Tickets Unclaimed_x	-0.037939
Non-prize remaining	-0.057584
Winning Tickets At Start_x	-0.089743
topprizeremain	-0.096124
Total at start	-0.132820
Change in Probability of Profit Prize	-0.175790
Change in Odds of Profit Prize	-0.208243
Rank by Cost	-0.223341
Non-prize at start	-0.232335
Rank by Best Probability of Winning Profit Prize	-0.293429
Odds of Profit Prize + 3 StdDevs	-0.321321
Rank by Best Probability of Winning Any Prize	-0.467783
Odds of Any Prize + 3 StdDevs	-0.513583
Odds of Profit Prize	-0.679773
Starting Odds of Profit Prize	-0.696712
Current Odds of Any Prize	-0.894417
overallodds	-0.894417

Name: Probability of Winning Any Prize, dtype: float64

Figure 1: Correlation Matrix

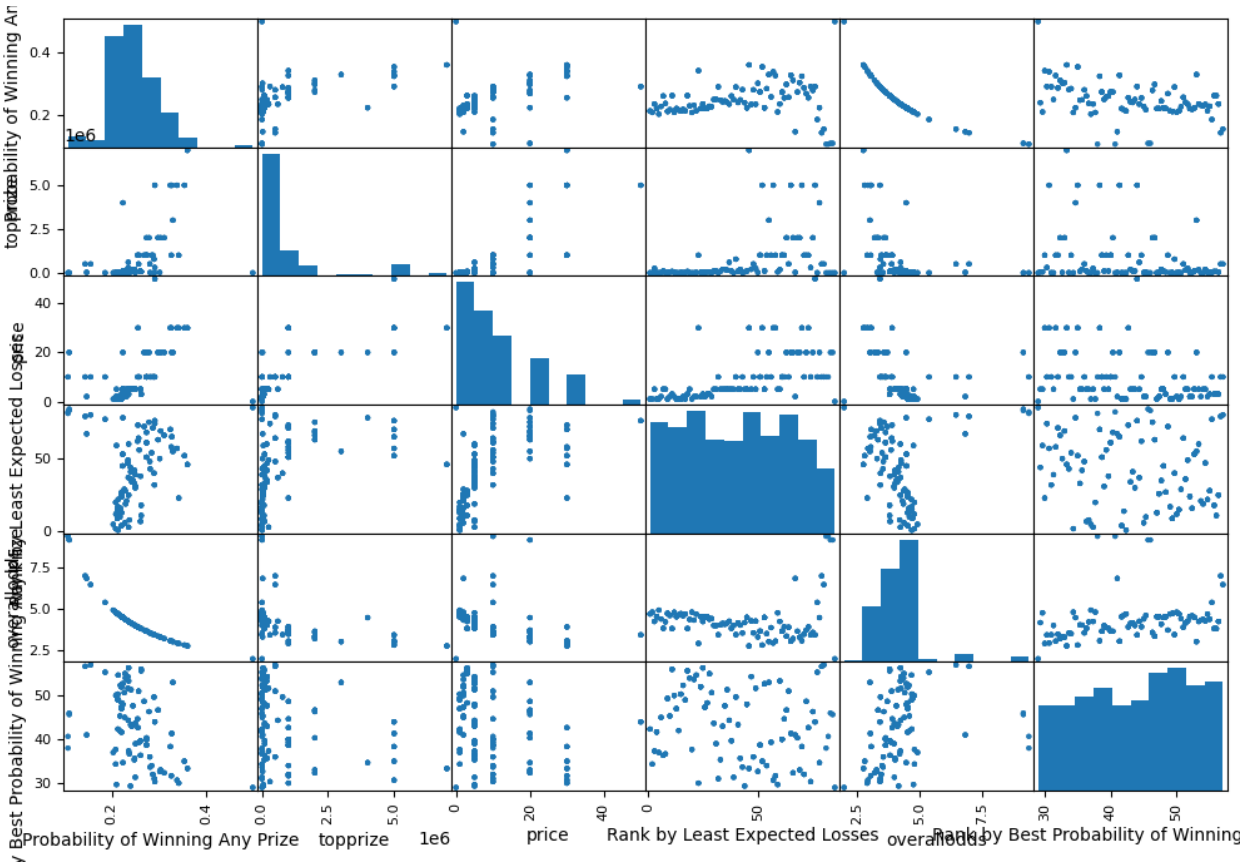

Figure 2: Scatter Matrix