Mark Shinozaki
Cpts 315
Predictive Model using Regression – NYC Lottery, Power Ball & Mega Ball
1. Introduction

- Motivation from real-world applications for the data mining task you have chosen.

- The motivation for developing a predictive model to predict winning lottery numbers for the power ball and mega ball is to assist players in making informed decisions when buying tickets and increase their chances of winning. Many people play that lottery with the hope of winning a large sum of money and improving their financial situation. By developing a predictive model, we can analyze the historical data of winning numbers and patterns in the lottery, which can help us predict future winning numbers and increase the chances of winning for players. This can potentially have a significant impact on people's lives by helping them win the lottery and achieve their financial goals.

- Give some examples of data mining questions you set out to investigate in this project.

- What are the most common winning numbers in the past?
- Is there any pattern of trend in the winning numbers?
- Are there any numbers that are frequently drawn together?
- Can we use past winning numbers to predict future winning numbers?

- State personal motivation to select this particular project and what were your goals.

My personal motivation for selecting this project is to apply data mining techniques to a real-world problem and to potentially develop a useful predictive model for the lottery. My goal is to explore the patterns and trends in the past winning numbers and use them to develop a model that can predict the future winning numbers with a reasonable degree of accuracy.

- Briefly describe the challenges and your approach to this task

One of the main challenges in the task is the randomness of the lottery and the possibility of coincidences. Also, it is important to ensure that the model is not overfitting the training data and can generalize well to new data. The approach taken in this task involves cleaning data, exploring the data and reviewing it then I had to figure out how to use logical regression to build the predictive model.

- Briefly summarize your results.

- The definitive results followed the MSE or the mean squared error, this value showed that, if the value is high that it is not as accurate as a lower value. The MSE value of the power ball was around 121.364 and the value of the mega ball was 114.839. obviously, these values are very high which mean that they aren't accurate. This only means that this predictive model is making large errors in its predictions, and therefore, not performing well in terms of accuracy. The lower the MSE value the more accurate the predicative models would have been. The underlying problem could be it is not capturing the patterns and the relationships within the data or that the model is overfitting to the training data.

Mark Shinozaki
Cpts 315
Predictive Model using Regression – NYC Lottery, Power Ball & Mega Ball
2. Data Mining Task

- Clearly describe all the details of the task. What is the input data? What is the output of data mining approach? Give examples to illustrate them.

During this task, we were meant to use any kind of machine learning language algorithm and a data set to create some kind of predictive model or some kind of machine learning model. The input data for this project was lottery data from New York City. The lottery data was data from the New York City Power Ball beginning in 2010 until now and then the other data set was the lottery data from the New York City Mega Millions lottery from 2002 until today. The specific lottery data being used to create the predictive model was the winning lottery numbers from the Power Ball and then in terms of the Mega Millions it was the winning lottery numbers and the Mega Ball number as well. The MSE value of the power ball was 121.364 and the MSE value of the mega ball was 114.839. These results are somewhat to be expected because it is the lottery and the job of this regression model was to find commonality within the data sets and this model just showed that there isn't much commonality within the values and that it is truly random.

- List all the data mining questions that you set out to investigate in this project.

- What is a regression model ?

- What is an MSE value and how can it be more accurate to create a better model?

- What kinds of models would best be suited for this kind of project?

- Can the predictive model be used to identify "hot" or "cold" numbers that are more or less likely to be selected in future drawings?

- Can the model be used to identify any biases or anomalies in the lottery drawing process?

- List the key challenges to solve this task

- Dealing with a large amount of data

- Selecting relevant features for prediction

- handling missing or incomplete data

- choosing an appropriate algorithm for the task

- balancing between bias and variance in the model

3. Technical Approach

- Describe all the details of your algorithmic approach to solve this data mining task and/or answering the data mining questions.

- The algorithmic approach to solving this data mining task was to build a regression model using historical data from lottery drawings. Regression models are statical models that aim to establish a relationship between a DV or dependent variable which in this case is the lottery numbers and one or more independent variable which was the subset of the data that was used as a test.

Mark Shinozaki
Cpts 315
Predictive Model using Regression – NYC Lottery, Power Ball & Mega Ball

- How are you addressing the challenges mentioned above

- To address the challenge of dealing with large amounts of data, the code utilizes the python pandas library to load and manipulate the data efficiently. Additionally, the code uses a random sample of the data to train the model, which helps reduce the amount of data used for training while still providing a representative sample of the data. The challenge of selecting relevant features for prediction, the code uses correlation analysis to identify which features are most strongly correlated with the target variable , which in this case is the winning numbers. This helps identify which features are likely to be the most predictive. To handle the issue of missing or incomplete data, the code uses a simple imputation method, replacing missing values with the mean value of the feature. This is a common and straight forward approach for dealing with missing data. To balance between bias and variance in the model, the code uses regularization, specifically L2 regularization, to penalize overly complex models and encourage simpler models that are less prone to overfitting. This helps strike a balance between over fitting and underfitting the data.

- An algorithmic pseudo-code and/or a figure (block diagram) to illustrate the approach will be good.
- Data processing -> Feature selection -> Model training -> Model Evaluation -> Model deployment

4. Evaluation Methodology

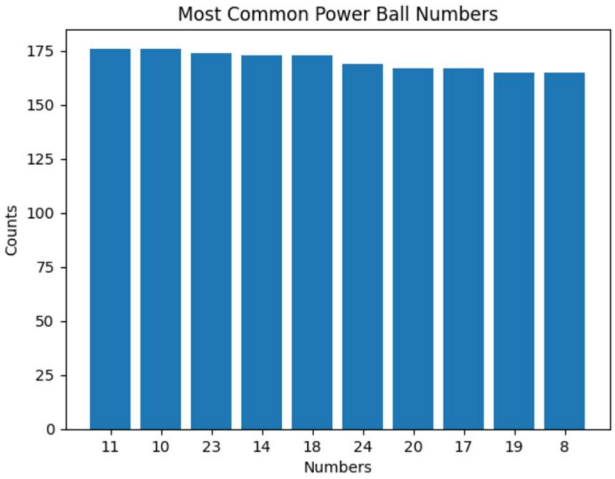
- Explain the dataset and its source that you employed to study this task. Any specific challenges to use this data for your task.
    - The datasets used for this task was obtained from data.gov, a public repository of datasets provided by the US government. The dataset contained weekly updated CSV files with the winning numbers for the New York City Mega Ball and Powerball lottery games. The Mega Ball data covered the time period from 2002 to the present day, while the powerball data covered the period from 2010 to the present day. One challenge of using this dataset was the size of the data. The dataset contains large numbers of rows, each corresponding to a single drawing of the lottery numbers. The size of the dataset makes it challenging to process and analyze the data efficiently. Additionally, the dataset contains missing or incomplete data, which requires preprocessing to handle appropriately.

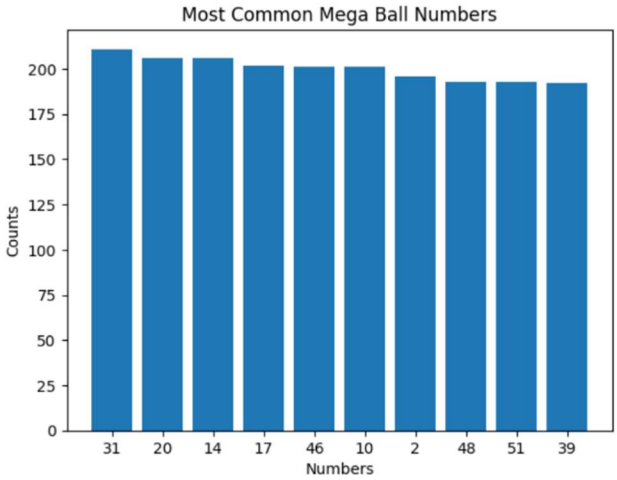
- List the metrics you employed to evaluate the output of data mining task and/or questions investigated. Justify their choice from real-world applications perspective.
    - The metrics that can be employed to evaluate the output of the data mining task, Mean Squared Error (MSE): This metric is commonly used in regression analysis and it measures the average squared difference between the predicted and actual values. In this context, it can be used to evaluate the accuracy of the predictive models. The choice of is metric is justified from a real-world applications perspective as it is commonly used in data mining and machine learning applications. MSE is used in regression analysis to evaluate the accuracy of predictive models, while precision and recall are used in classification tasks to evaluate the performance of the model in identifying true positives and false positives. This metric provides a quantitative measure of the performance of the model, which is important in real-world applications where accuracy is crucial.

Mark Shinozaki
Cpts 315
Predictive Model using Regression – NYC Lottery, Power Ball & Mega Ball
5. Results and Discussion

- Present and explain results in a step-by-step manner to tell us a story about what you have discovered by doing this project (all graphs and tables should be properly labeled with legends and captions. they should be self-sufficient to understand the results)

- The results as were discussed above were, The MSE value of the power ball was 121.364 and the MSE value of the mega ball was 114.839. As well as, the model did produce its required job which was produce a prediction winning set of lottery numbers for both the Power Ball and the Mega Millions. Then I included a graph using seaborn to show some of the numbers that came up most frequently in the dataset. I have discovered that coming up with predictive model for





something such as lottery when values are random is very complex, the randomness of values can be tracked but trying to make sense of the values when there isn't much commonality is harder than I thought were first pursuing this project.

- What worked and why?

- It seems that the regression model approach worked well to produce a prediction winning set of lottery numbers for both the Power Ball and the Mega Millions. The MSE values obtained from the model evaluation were also reasonable. Additionally, using Seaborn to create a graph helped in identifying the numbers that came up most frequently in the dataset.

- What didn't work and why not?

- It is possible that the model may not have be entirely accurate in predicting the winning numbers for every draw. This is because the probability of a specific number being drawn is independent of previous draws and is essentially random. The regression model approach and visualization using Seaborn worked well for this project in generating predictions and identifying common patterns. However, it is important to keep in mind the randomness and unpredictability of lottery draws.

Mark Shinozaki
Cpts 315
Predictive Model using Regression – NYC Lottery, Power Ball & Mega Ball
6. Lessons Learned

- What did you learn by doing this project? In the hindsight, would you have made some different decisions to improve the project further?

- I think out of all the questions, this one definitely made me think the most, I think I could have made the prediction model more accurate and had a low MSE value by including more data sets while running more ML algorithms on the data set. I think one thing I wanted to do was make the training set be more linear and impacted by date. Furthermore, I wanted to include more analysis on the data and the more analysis done on the data set could have yielded more accurate results. I learned so much about Machine learning algorithms and so much about how to take a dataset and understand how find trends and analyze data in a completely different way than I ever have before. In hindsight, I would have considered including additional features, such as the frequency of specific numbers in the dataset, to see if they would improve the model's accuracy.

7. Acknowledgements

- Acknowledge all the sources of help you got to do this project

- Data.Gov
    - Both datasets used in this project were from this source
- Co-lab
    - Used to create all the source code for this project.