



THE LOTTERY LOT

Natsuura Fukuda

DO YOU HAVE WHAT IT
TAKES TO WIN A MILLION
DOLLARS? (OR TEN)



Overview

Ever wondered if the lottery has some pattern or bias?

What if it's possible to see who's more likely to win by analyzing winners's demographic data?

By aggregating information about Texas Lottery winners we can pinpoint aspects that make those more fortunate more likely to win.

By uncovering these biases, the lottery can become more fair and equal.



1

2

3

4

5

6

7

The Data



1

The dataset used “Winners List of Texas Lottery Prizes” which lists location of winners and retailers, cost of tickets, date of purchase, and type of ticket.

2

There’s over 2 million data points making it difficult to compile due to runtime problems.

3

4

5






By cutting data that was negligent (like if the winner chose annuity or not) I reduced the total amount of data.

6

7

What's in this Dataset?

Columns (35)

Column Name	Description	API Field Name	Data Type
Tr Row ID	Unique key.	row_id	Text
 Claim Number	A number assigned by the Claims Process. A claim may contain multiple claim lines.	claim_number	Number
# Amount Won	Amount won for winning play if cash prize. Prize value for 2nd Chance promotion if merchandise prize.	won_amount	Number
 Date Claim Paid	The date the claim was paid.	claim_paid_date	Floating Timestamp
 Player ID	Number assigned by the payment processing system to the claimant, or claim number for anonymous claimant.	player_id	Number
 Annuity Indicator	Indicator if the claim is an annuity (YES) or cash value option (NO).	annuity_indicator	Text
 Anonymous Indicator	Indicator of player's request to remain anonymous.	anonymity_indicator	Text

Each row is a
**Winning claim,
player, prize,
game and
selling retailer
identity and
location
information.**

Rows
2.87M

Columns
35

Row Identifier
Row ID

The Data (2)

1

2

3

4

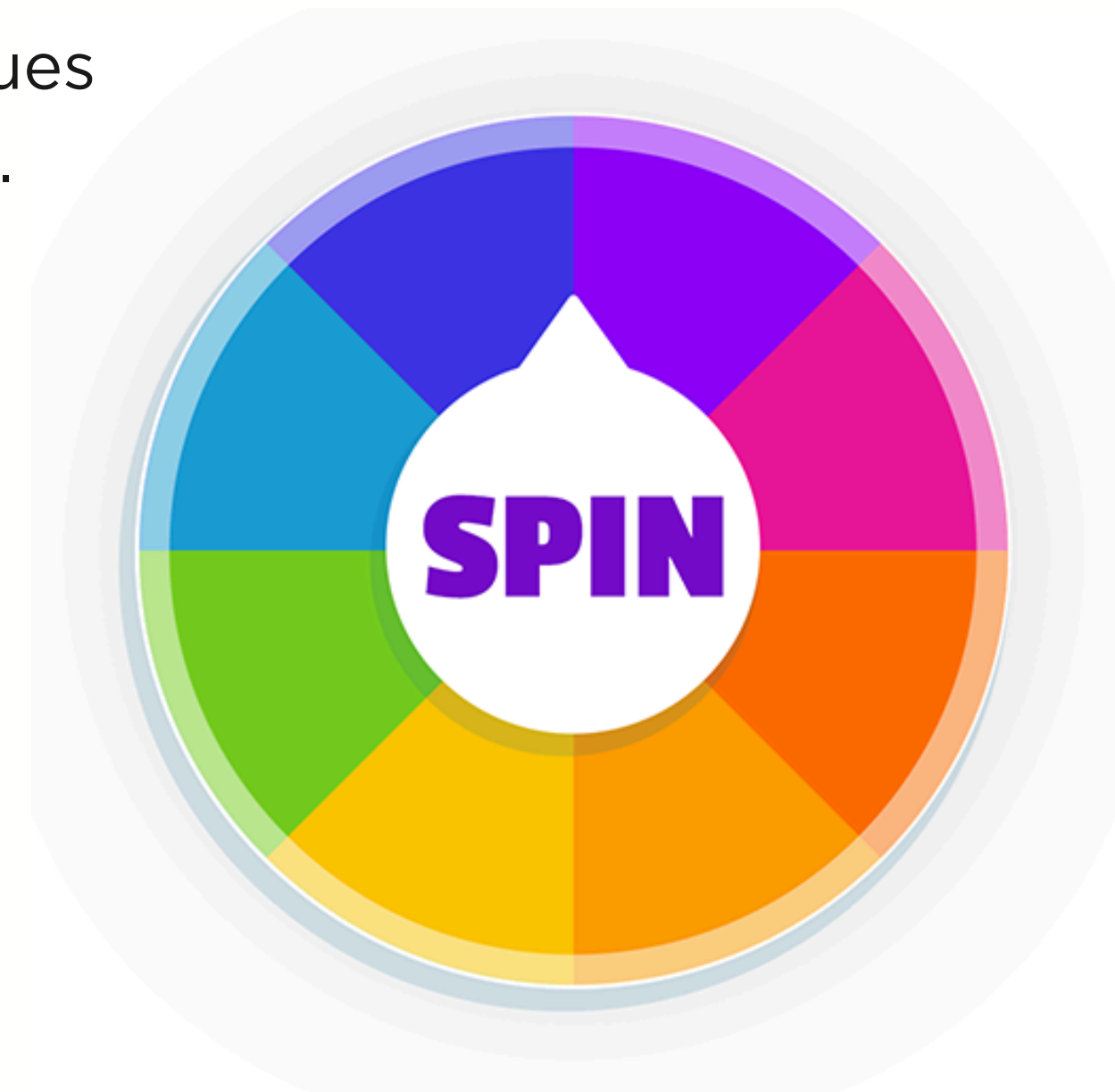
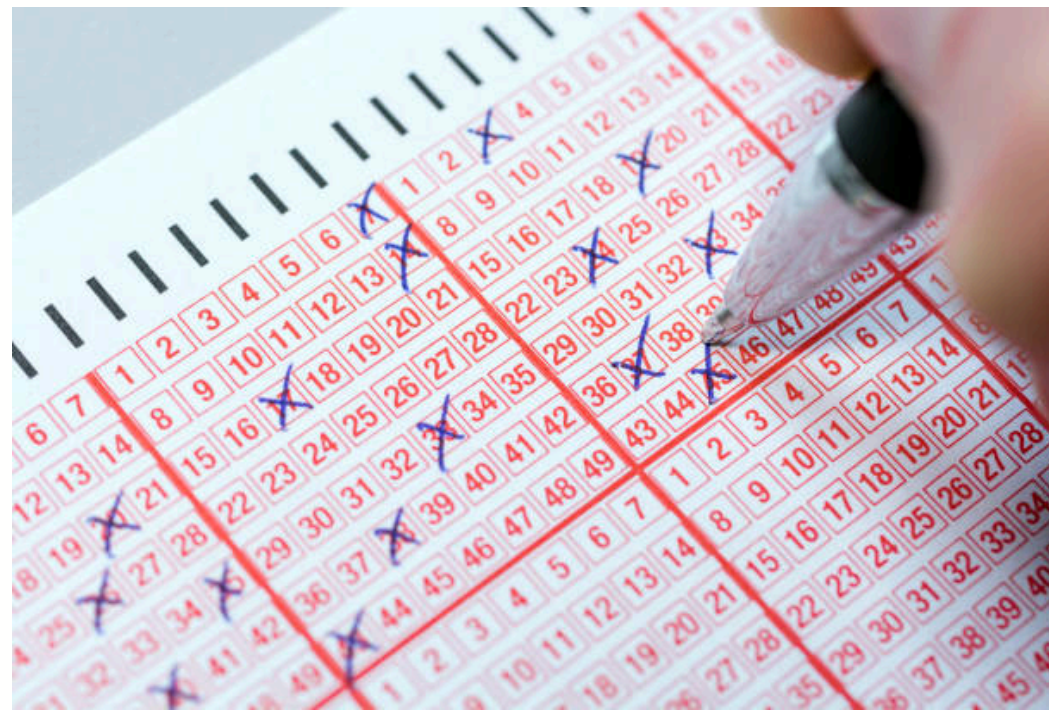
5

6

7

I divided the dataset into two groups, scratch tickets and draw tickets since the former had scratch ticket cost and the latter had draw game type which the other didn't have.

There were many empty data cells so I used random values based on the proportion of existing values to fill them in.



Preliminary Findings

1

I used a regression decision tree to set a baseline model for my data but due to the size of the data, scaling issues, and class imbalances.

2

3

I limited data to the 100 most prevalent categories in each class to make it easier to graph, run, and understand.

4

5

I encoded my data using frequency encoding since it's a very large dataset and most of my classes were categorical.

6

7

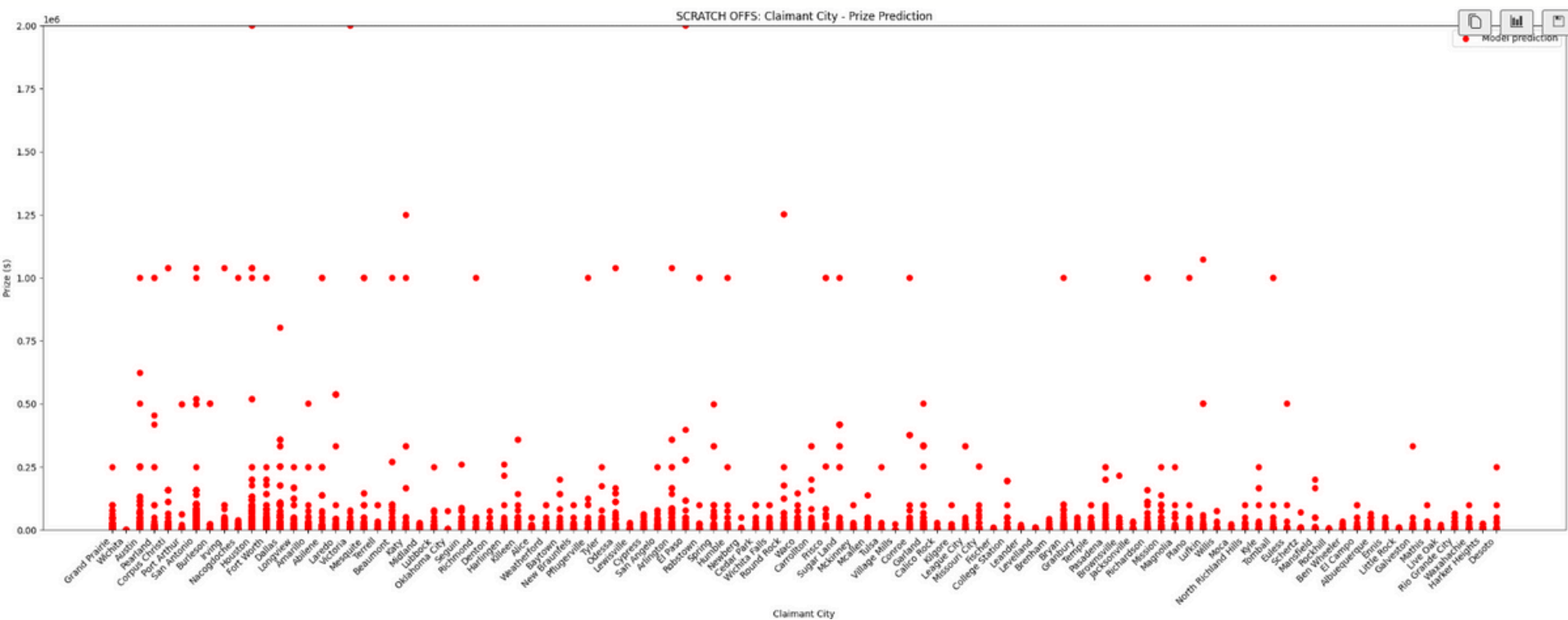
**The MSE (Mean Squared Error) was
~6.8 billion for scratch tickets
and ~920 billion for draw tickets.**

**The R-Squared was
-0.59 for scratch tickets
and -0.40 for draw tickets.**

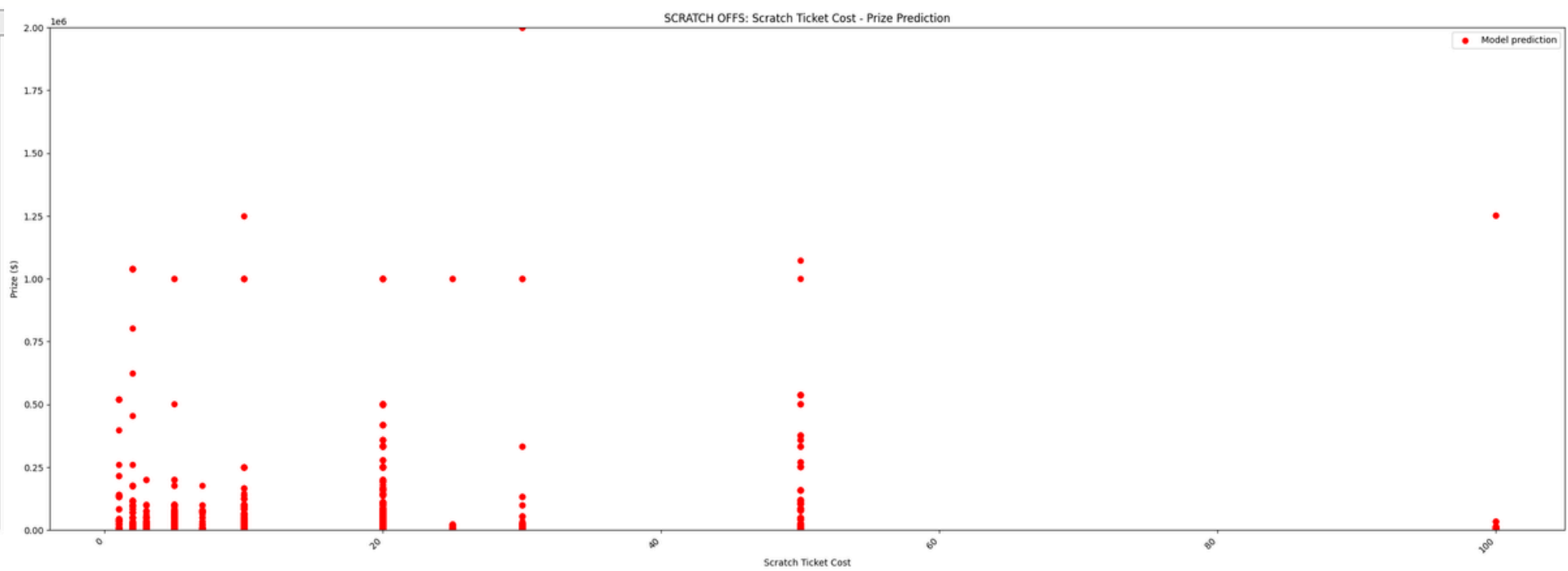
Not good...

Figures from Baseline Modeling

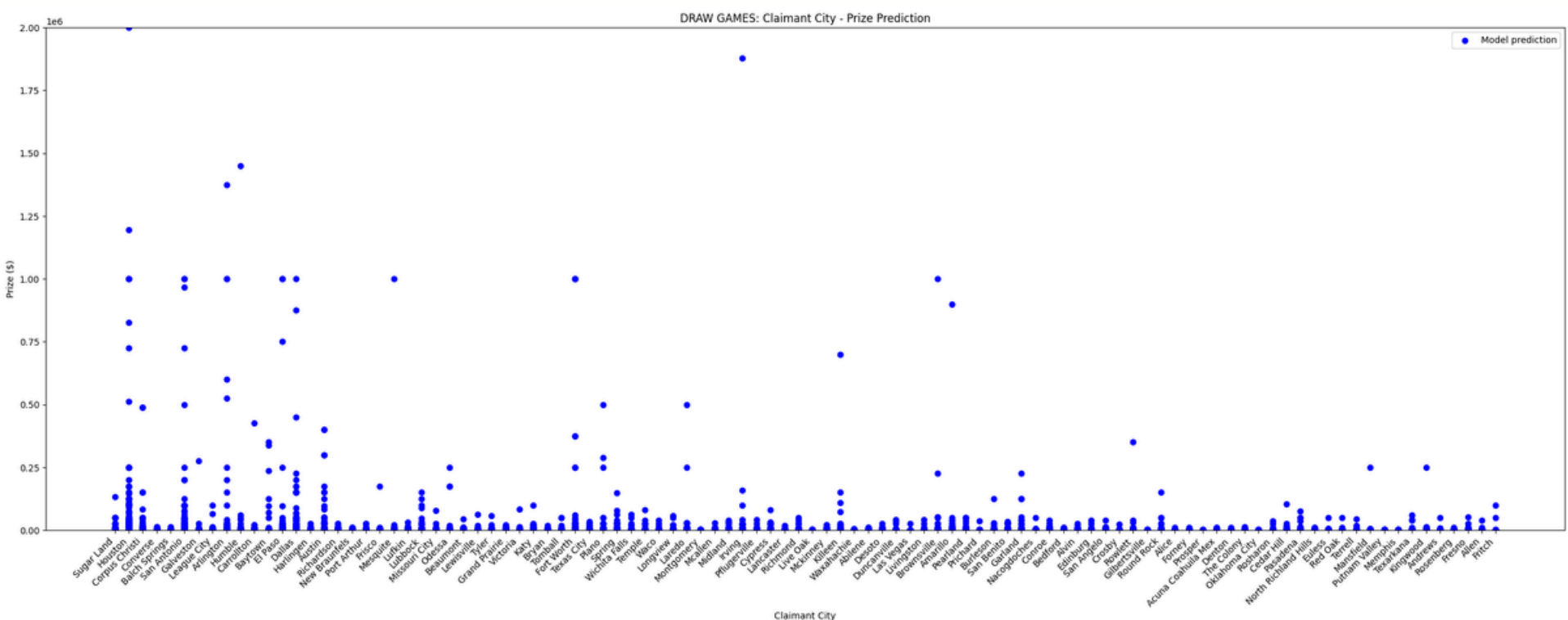
Claimant City to Predicted Prize (Scratch Tickets)



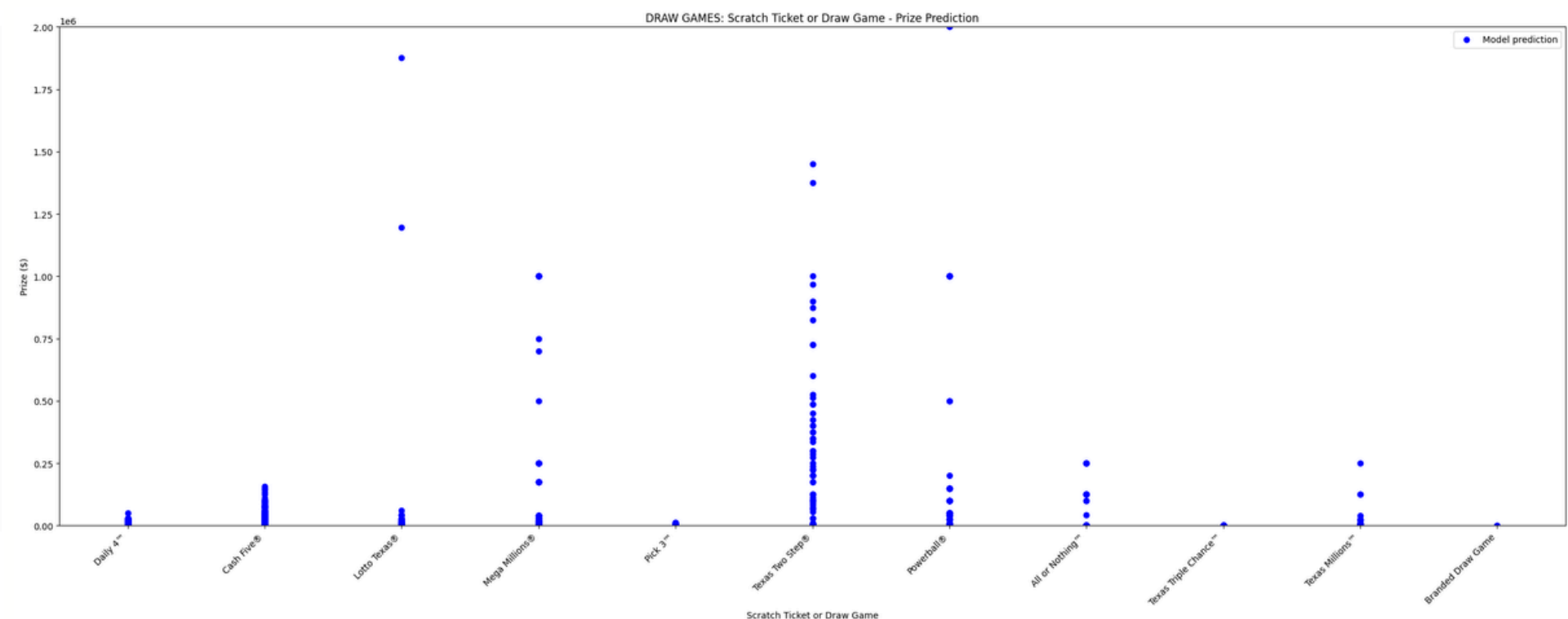
Ticket Price to Predicted Prize (Scratch Tickets)



Claimant City to Predicted Prize (Draw Tickets)



Draw Game Type to Predicted Prize (Draw Tickets)



Final Model

1

2

3

4

5

6

7

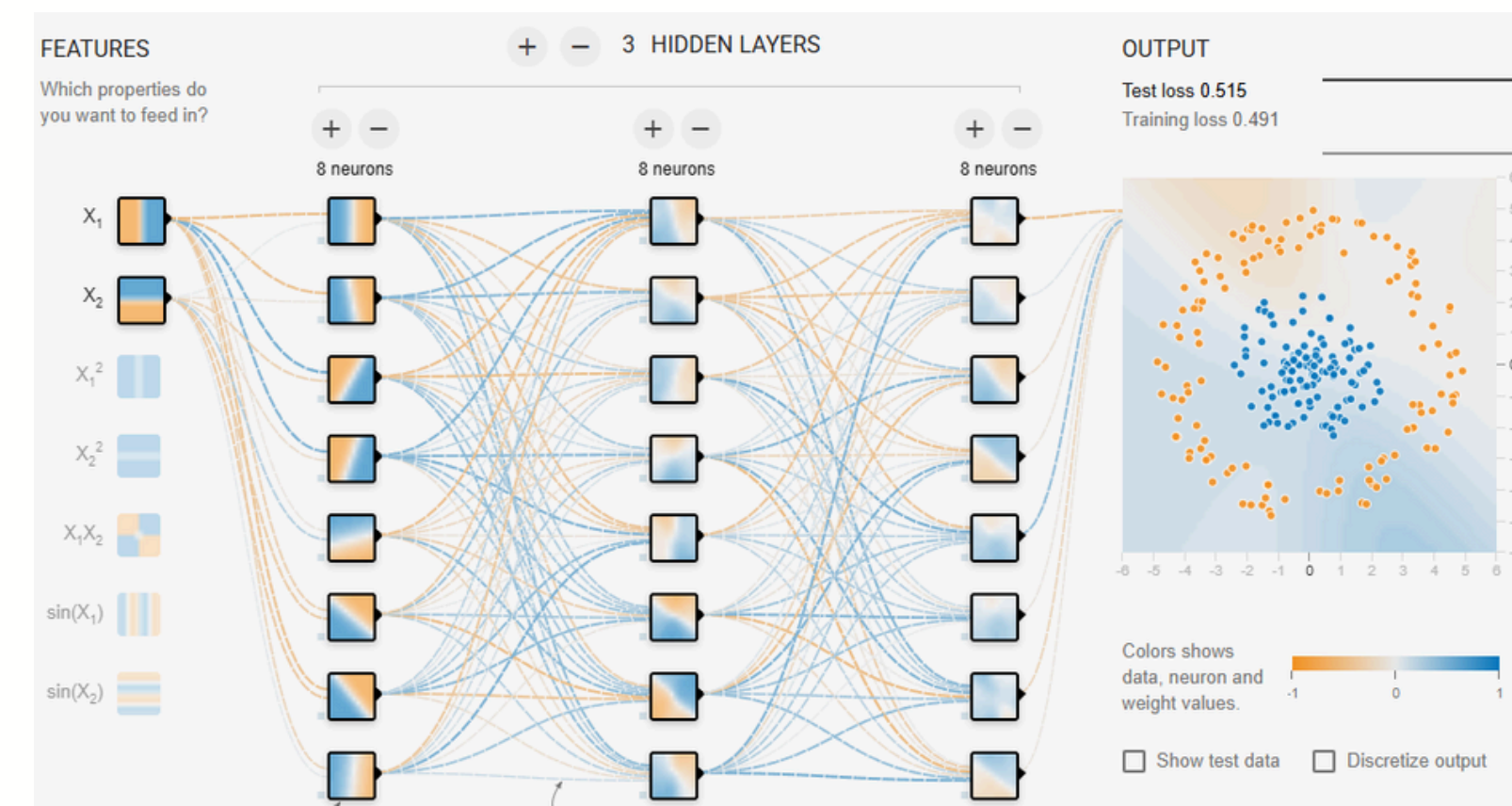
I chose a simple neural network for my final model

I improved on the baseline by putting the prize values (y-value/target) through a natural log function, standardized the data (x-value) by scaling it, and sending it through a 3-layer 25 node network.

By putting the y-value into a log function, it reduced the range that used to be from 1 - 400 million to 1 - 20, reducing outliers.

By standardized data to make each class have the same scale.

This but each layer has 25 instead of 8



Final Model (2)

1

2

3

4

5

6

7

By interpreting the final model we get that being:

A white person from/living in a state in Mexico (specifically Nuevo Leon, Aguascalientes, or Chihuahua) and a small city buying a \$30 scratch off ticket or a lower prize pool draw game like Cash Five or Daily 4 from Kroger or Winner's Corner is more likely to yield higher lottery winnings

**The MSE (Mean Squared Error) was
~3.47 for scratch tickets
and ~3.21 for draw tickets.**

**The R-Squared was
0.56 for scratch tickets
and 0.45 for draw tickets.**

The final model fits much better than the baseline model

Notable characteristics for high prizes are that

- 4 out of 6 top Claimer Cities are mid-populated (around 5,000-30,000) and are mostly white even when Texas is 39% white and hispanic each
- All the highest Claimer States for scratch tickets are in Mexico
- \$30 scratch tickets yielded the highest average prizes even without taking into consideration the profit made, outdoing \$100 and \$50 tickets
- 2 of the 3 highest prize draw games have relatively low jackpots in the \$5,000 to \$25,000 range
- Kroger was one of the top prized retailers in both scratch and draw tickets
- Winner's Corner is a lottery retailer, specifically selling lottery tickets

Notable characteristics for low prizes are that

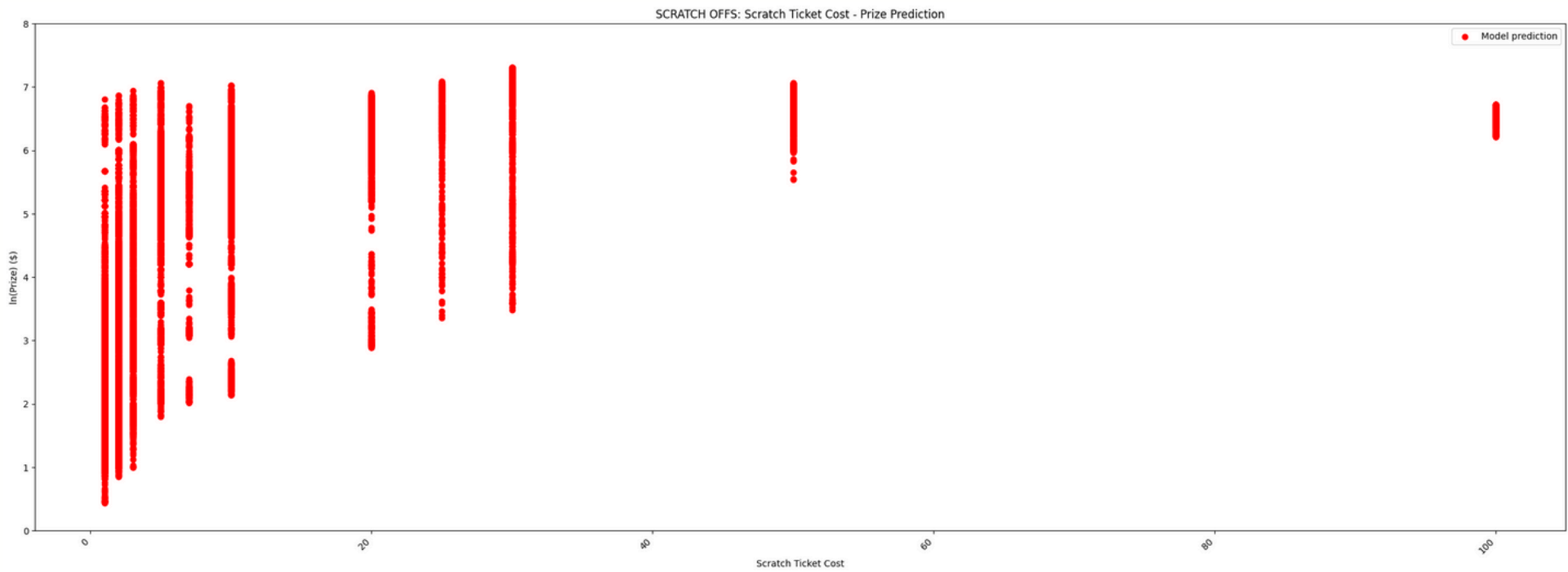
- Oklahoma City is in the top 3 lowest prized Claimer Cities in both scratch and draw games
- Gilbertsville has a population of only 399 in the 2010 census and is located in New York
- None of the lowest prized Claimer Cities are located in Texas
- El Paso county/city was in the top 3 lowest prized for Claimer County and Retailer County for draw games, and Retailer City for both types
- A majority of the lowest prized retailers specialize in food/supermarkets
- El Paso and Oklahoma City have similar large populations of around 600,000

Figures from Baseline Modeling

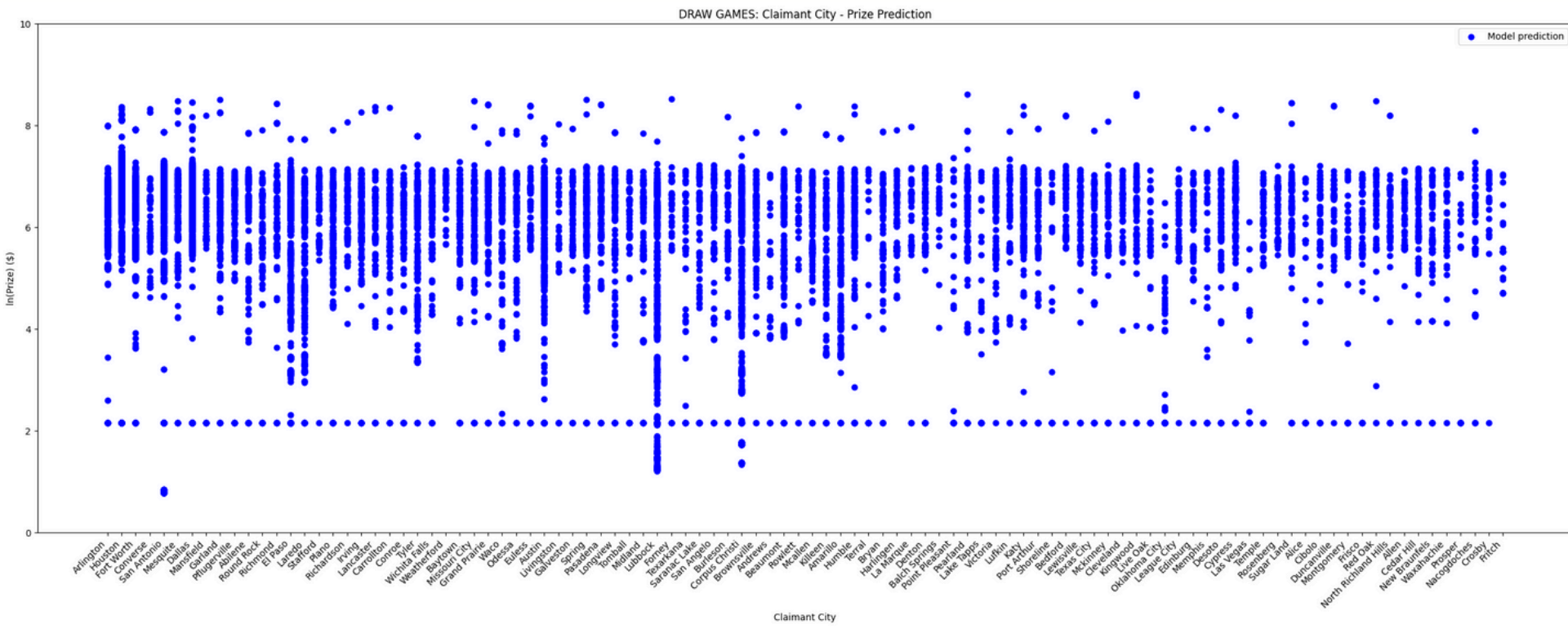
Claimant City to Predicted Prize (Scratch Tickets)



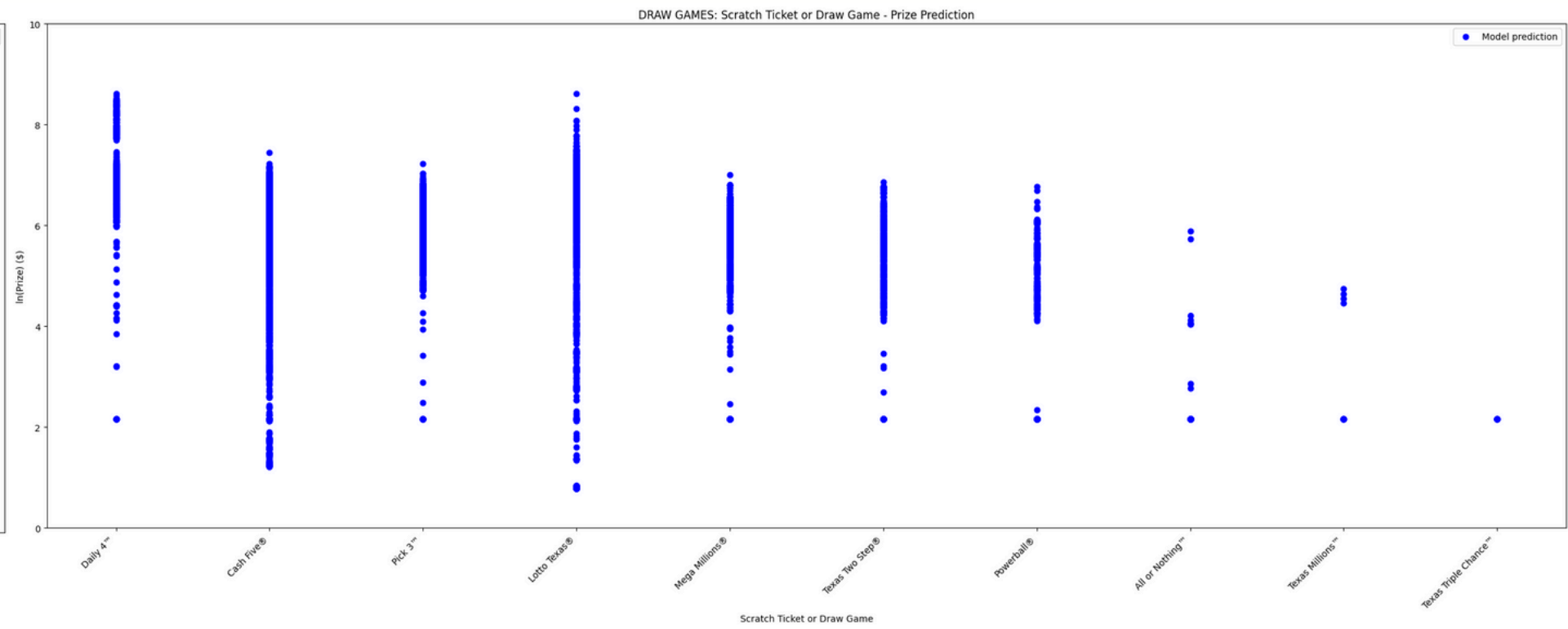
Ticket Price to Predicted Prize (Scratch Tickets)



Claimant City to Predicted Prize (Draw Tickets)



Draw Game Type to Predicted Prize (Draw Tickets)



Next Steps

1

I would like to gather data on all lottery winners in the US, not just limited to the Texas lottery to get a more comprehensive view of the whole problem.

2

3

Not only gathering data, but adding categories like age, gender, and income would greatly help to identify specific demographics.

4

5

If the runtime and machine permits, I would like to increase complexity of the neural network as the only limiter to not increasing complexity was the runtime.

6

7





**Thank
you**

