CS542 Common Task - Simon Kye

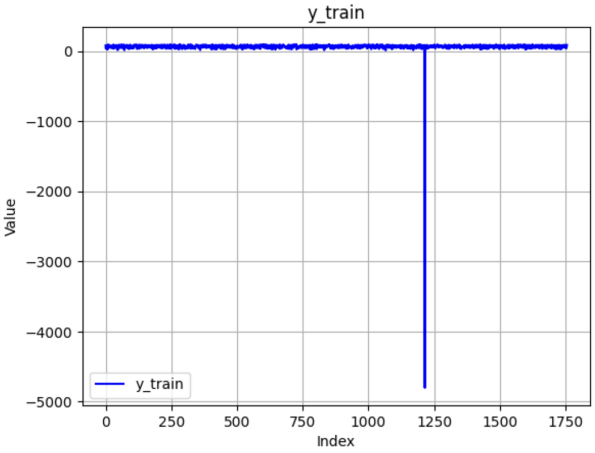
This PDF is paired with a IPNYB in the Jupyter Notebook folder. i.e. When there is [Cellblock 1], look at the IPYNB file and look for the cell block under the text labeled “Cellblock 1”. Majority of the graphing was done with the use ChatGPT.

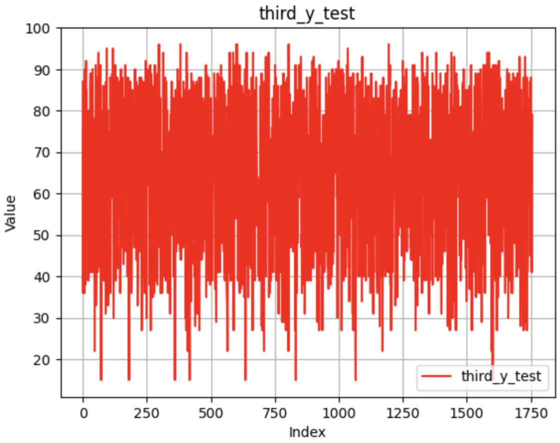
1. Introduction

The goal of this common task was to predict the max temperature of New York for the day. We aimed to forecast the max temperature through linear regression. The dataset and the type of dataset used goes through multiple iterations as we improve the dataset used to ensure the most accurate results.

2. Dataset

From Sep 25-26, we use daily information of 2016 from ERA-5, JFK Airport, VisualCrossing API, and WeatherStack API, and a Kaggle Central Park data [Cellblock 1]. After being informed in lecture that our dataset should be hourly, we switched from the above sources to hourly data of 2016 from ERA-5, JFK Airport, LaGuardia Airport, NOAA’s Manhattan data, and Buffalo Niagara International Airport [Cellblock 2]. Finally, after realizing that the data used for Manhattan is of Kansas, Manhattan, we switch to Ithaca Tompkins International Airport [Cellblock 3]. It’s important to note that all sources listed here resides in New York. For missing data points in the CSV’s downloaded from some of these sources, we cleaned up the data by looking at the closest previous day and closest future day and interpolating the datapoints. After checking the cost functions, we see the third model has a very high MSE and R^2. We find the outlier by checking values of the training y values and testing y values. We see there is an extreme outlier and remove the datapoint in the dataset that is responsible for this. [Cellblock 4 & 5].





3. Model

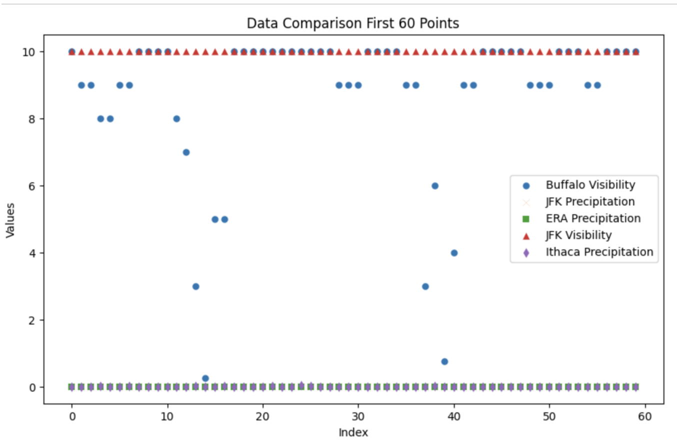
For this task, we ended up using linear regression for the task. In order to prevent overfitting on this small dataset, we generalize using ridge regression [Cellblock 11]. There are many better models that could've been used. But it is important to note that due to unfamiliarity with coding machine learning, it was very difficult to implement more suitable models such as Autoregressive Integrated Moving Average (ARIMA) or Seasonal Decomposition of Time Series (STL) as they're not in SkLearn and have less resources on them. Despite this, the modifications made to the data should still be applicable and helpful to any other model used.

4. Training

From our first model using daily dataset, training was done by setting X as "yesterday"'s weather parameters in the different locations, whether that be precipitation, temperature, humidity, etc. We then set Y as the high temperature in central park for "today" [Cellblock 1]. We can then use this trained model to input yesterday's data points to predict today's high temperature. The second data set, which is hourly, uses a similar method [Cellblock 2]. Rather than yesterday's data, it trains off previous hour's data to predict the max temperature for the day. It takes into account which hour of the day it is to adjust the slope accordingly. This is then further improved upon by taking in the previous 24 hours as the input to predict the current day's max temperature [Cellblock 3]. We set our cost function as MSE as it is a regression problem.

5. Features

In the results sections, we'll see that not all features are beneficial to the training of the model. Most notable of these being precipitation and visibility. For almost all days, it is recorded that there is 0 inches of precipitation and 10 miles of visibility. We can see from a visualization sampling the first 60 days of visibility and precipitation that this is the case.



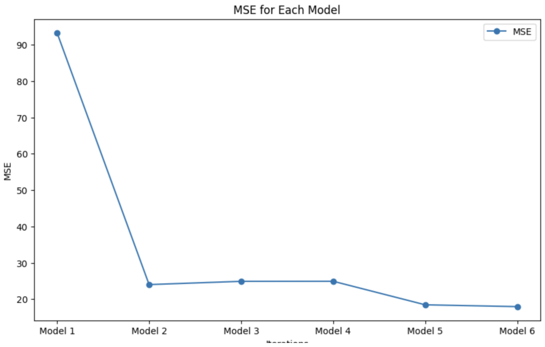
Because of this, the weights for these features are comparatively high to other features. When there is a day where it rains or when there is a day below 10 miles of visibility, it usually affected the temperature more than it should've. Due to this, precipitation and visibility was removed moving on from 2nd to 3rd dataset, which lowered the MSE from 24.92 to 23.18.

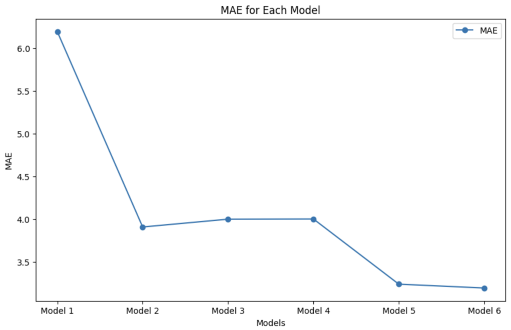
6. **Purchasing**

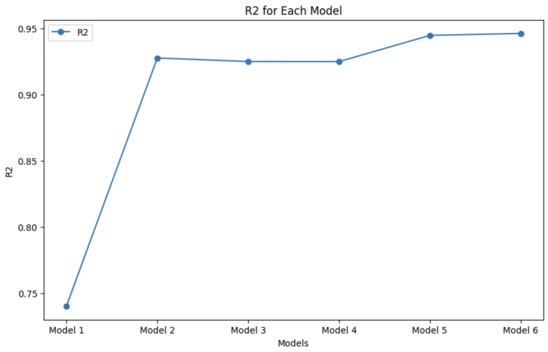
In order to remove as many dependent variables as possible, we decided to stick to always buying the same amount of 'Yes' stock for the same price at the same hour [Cellblock 12]. The hour was determined by splitting the predictions based on which it hour predicted in [Cellblock 10]. We chose hour 10, as although it wasn't the minimum, it was relatively small and was early enough in the day to lead to a noticeable profit. No matter how unrealistic the prediction looked, we would always purchase the stock. This was to isolate the model's ability to predict the high temp of New York accurately and nothing else.

7. **Results**

We see that through the refinement of the dataset and how we use the dataset, we all performance metrics such as MSE and MAE goes down and R^2 goes up [Cellblock 15]. Although the second model "performed" better compared to third and fourth model, it was trained off Kansas, Manhattan and shouldn't be regarded.

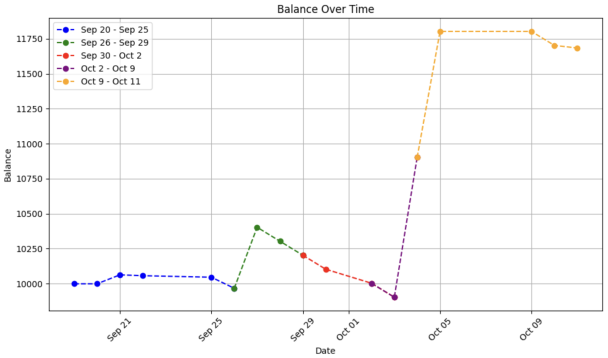


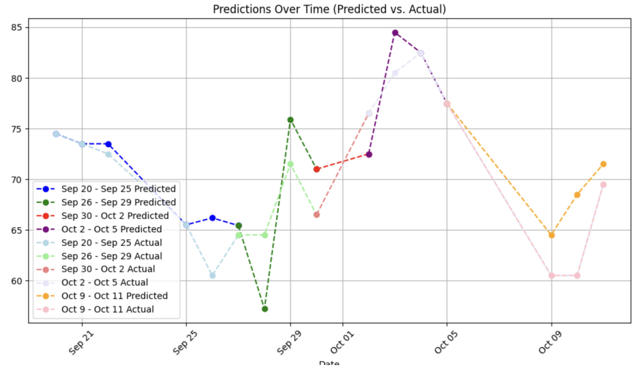




We make improvements to our predictions through various methods, but the most notable of this was the number of datasets. By increasing our data points from 365 to 8760 data points, the model was able to be much more accurate. We also see that providing more context for the input increased its accuracy as well. Instead of simply putting in the previous hour's data, we input the last 24 hours that occurred before the current hour.

Observe the graphs below showing balance over time and predictions over time [Cellblock 13 & 14]. The blue lines are manual trades, where we used the weather forecast. We see that although it was very accurate, very little profit was made due to the predictability of the result. Green was with the incorrect dataset. Red was with previous hour data, and purple was with predictions with the past 24 hours. There are too few data points to make any conclusions, but it is interesting that through these predictions and sticking to these predictions no matter what, we're able to make much higher profits by not inherently selecting the predictable answer.





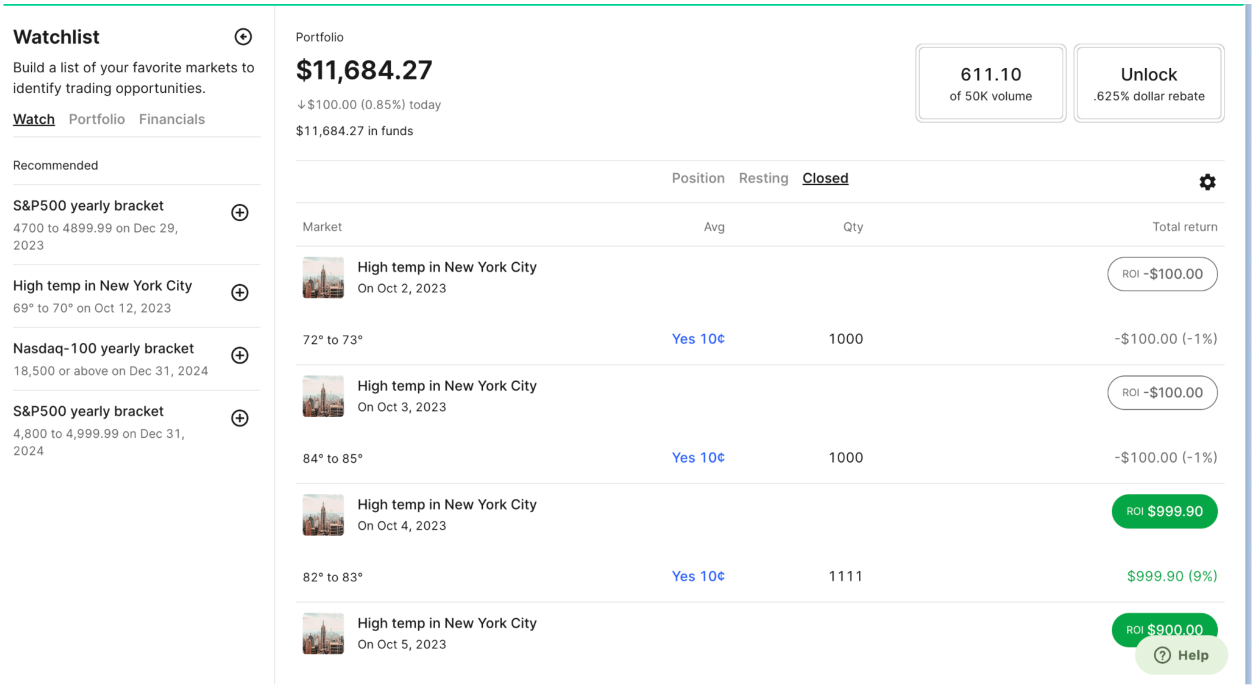
We also see that the model is better than a random prediction based off our current data. Since there are six different options in Kalshi, if we were to pick randomly, we would get the correct answer ⅙ times. But we see that after starting to use the various models, we got the correct answer 5/15 times.

8. Reflection & Conclusion

Through the common task I learned some very crucial things about machine learning. Mainly being the impact of the quality and the size of the dataset we use. I've come to realize that the quality of data is paramount in machine learning. Clean, accurate, and well-labeled data not only enhances the model's performance but also minimizes the risk of misleading predictions and is a sizeable chunk of the workload when it comes to machine learning. Data preprocessing techniques, such as handling missing values and outliers, have become clear necessities to ensure data integrity. The dataset's size also had a direct impact on the model's ability to generalize and make accurate predictions. Larger datasets often lead to more reliable models. Finally, I learned machine learning is not a one-time process but an iterative one that requires trial and error. It involves continuous learning, experimentation, and refinement. In conclusion, I believe this task gave a very surprising amount of insight as to how machine learning operates.

9. Kalshi Trade History

The screenshots here show the result of the Kalshi trades and is proof to the claims made above.



Watchlist

Build a list of your favorite markets to identify trading opportunities.

Recommended

S&P500 yearly bracket	4700 to 4899.99 on Dec 29, 2023
High temp in New York City	69° to 70° on Oct 12, 2023
Nasdaq-100 yearly bracket	18,500 or above on Dec 31, 2024
S&P500 yearly bracket	4,800 to 4,999.99 on Dec 31, 2024

Portfolio

$11,684.27

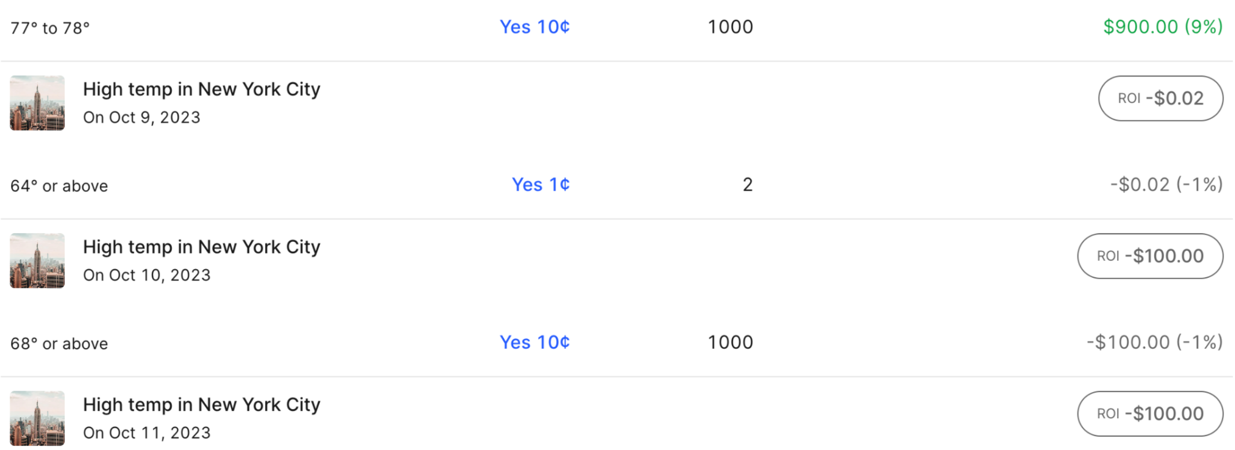
↓$100.00 (0.85%) today

$11,684.27 in funds

611.10
of 50K volume

Unlock
.625% dollar rebate

Market	Avg	Qty	Total return
High temp in New York City On Oct 2, 2023			ROI -$100.00
72° to 73°	Yes 10¢	1000	-$100.00 (-1%)
High temp in New York City On Oct 3, 2023			ROI -$100.00
84° to 85°	Yes 10¢	1000	-$100.00 (-1%)
High temp in New York City On Oct 4, 2023			ROI $999.90
82° to 83°	Yes 10¢	1111	$999.90 (9%)
High temp in New York City On Oct 5, 2023			ROI $900.00



77° to 78°	Yes 10¢	1000	$900.00 (9%)
High temp in New York City On Oct 9, 2023			ROI -$0.02
64° or above	Yes 1¢	2	-$0.02 (-1%)
High temp in New York City On Oct 10, 2023			ROI -$100.00
68° or above	Yes 10¢	1000	-$100.00 (-1%)
High temp in New York City On Oct 11, 2023			ROI -$100.00

High temp in New York City On Sep 21, 2023			ROI $63.90
73° to 74°	No 85¢	426	$63.90 (0.18%)
High temp in New York City On Sep 22, 2023			ROI -$6.00
73° to 74°	Yes 6¢	100	-$6.00 (-1%)
High temp in New York City On Sep 25, 2023			ROI -$11.00
65° to 66°	No 11¢	100	-$11.00 (-1%)
High temp in New York City On Sep 26, 2023			ROI -$96.04
66° or above	Yes 10¢	1000	-$96.04 (-1%)
High temp in New York City On Sep 27, 2023			ROI $436.19
64° to 65°	Yes 47¢	823	$436.19 (1.13%)
High temp in New York City On Sep 28, 2023			ROI -$100.00
61° or below	Yes 10¢	1000	-$1
High temp in New York City On Sep 30, 2023			ROI -$100.00
70° or above	Yes 10¢	1000	-$100.00 (-1%)

-city#highny-23sep21

Help