Philadelphia Crime

Nicholas Morris

Machine Learning

Wrangling → Feature Engineering → Modeling

Weekly Crime Rate

Computing weekly crime rate by district. Adding previous weeks of crime to the data.

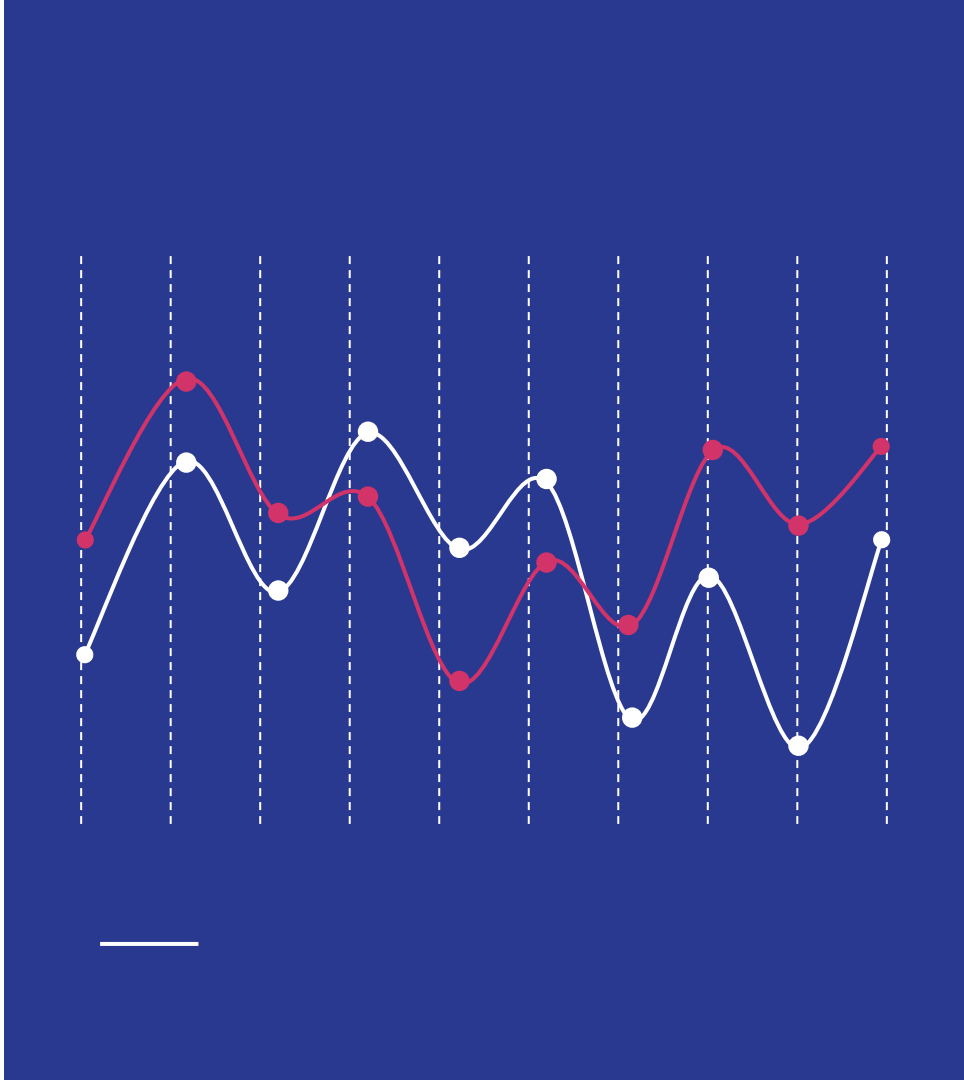
Additional Data

Adding economic trends such as unemployment and CPI. Converting timestamp components and district to binary data points. Attempting feature transformations.

Predictions

Training linear regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

Below is the first and last arrest in the data. There are 2,237,605 total arrests and 14 columns. The key columns are Dispatch_Date and Dc_Dist telling us when and where arrests occur. [Link to the dataset and code]

Dc_Dist	Psa	Dispatch_Date_Time	Dispatch_Date	Dispatch_Time	Hour	Dc_Key	Location_Block	UCR_General	Text_General_Code	Police_Districts	Month	Lon	Lat
3	I	2006-01-01 00:00:00	2006-01-01	00:00:00	0	200603000002	S 8TH ST /SOUTH ST	2600.0	All Other Offenses	3.0	2006-01	-75.155491	39.942416
24	3	2017-03-23 01:29:00	2017-03-23	01:29:00	1	201724026395	3700 BLOCK RICHMOND ST	400.0	Aggravated Assault No Firearm	17.0	2017-03	-75.087735	39.990 09

Crime Rate

Dispatch_Date	Dc_Dist	Crimes
2006-01-01	1	14.0
2006-01-02	1	17.0
2017-03-22	92	0.0
2017-03-23	92	0.0

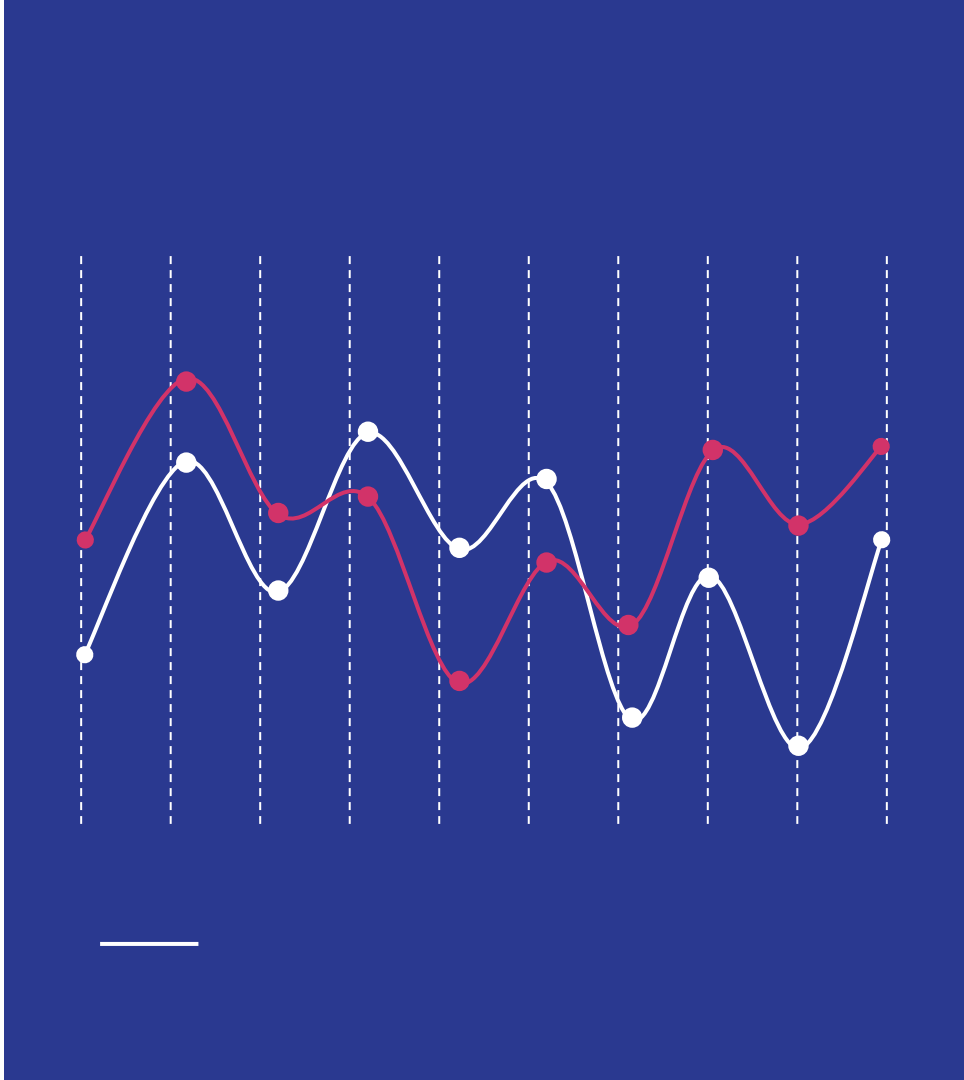
Aggregating the data by Dispatch_Date and Dc_Dist we get 102,500 rows of crime data.

Previous Weeks Of Crime

Aggregating the data by Year, Week, and Dc_Dist we get 14,575 rows of crime data. For each district the previous four weeks of crime were extracted.

Dc_Dist	Year	Week	Crimes	Crimes(t-1)	Crimes(t-2)	Crimes(t-3)	Crimes(t-4)
1	2006	4	97.0	144.0	103.0	129.0	14.0
1	2006	5	110.0	97.0	144.0	103.0	129.0
1	2006	6	77.0	110.0	97.0	144.0	103.0
1	2006	7	83.0	77.0	110.0	97.0	144.0

Feature Engineering



Time Components

Year	Quarter	Month	Week	Days	Crimes
2006	1	1	4	7	97.0
2006	1	1	5	7	110.0
2006	1	2	6	7	77.0
2006	1	2	7	7	83.0

The Year, Quarter, Month, Week, and number of Days in a week were extracted from Dispatch_Date.

Economic Data

The NASDAQ closing price, Unemployment rate, CPI, PPI, GDP, GDI, and Federal Funds Rate were pulled from FRED (Federal Reserve Economic Data).

Year	Week	NASDAQ	Unemployment	CPI	PPI	GDP	GDI	Federal_Funds_Rate
2006	4	2281.437143	4.700000	0.762195	164.300000	13599.16	13795.893	4.290000
2006	5	2284.638571	4.771429	0.361852	162.514286	13599.16	13795.893	4.432857
2006	6	2258.892857	4.800000	0.201715	161.800000	13599.16	13795.893	4.490000
2006	7	2274.302857	4.800000	0.201715	161.800000	13599.16	13795.893	4.490000

Binary Data

The time components and Dc_Dist were converted to binary variables.

Year_2006	…	Week_4	Week_5	…	Dc_Dist_1	…	Dc_Dist_92	Crimes
1	…	1	0	…	1	…	0	97.0
1	…	0	1	…	1	…	0	110.0
1	…	0	0	…	1	…	0	77.0
1	…	0	0	…	1	…	0	83.0

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

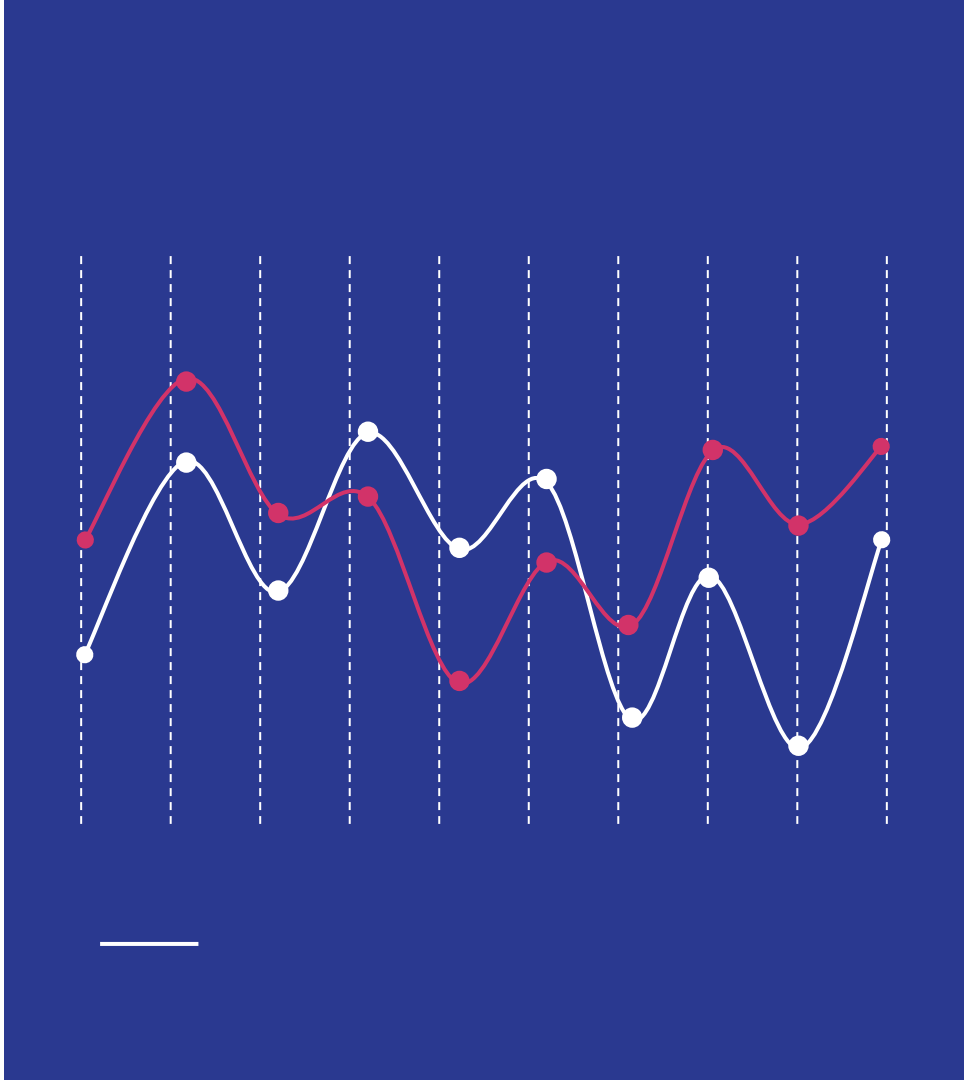
Reciprocals did not improve model performance, so, it was left out of the final model.

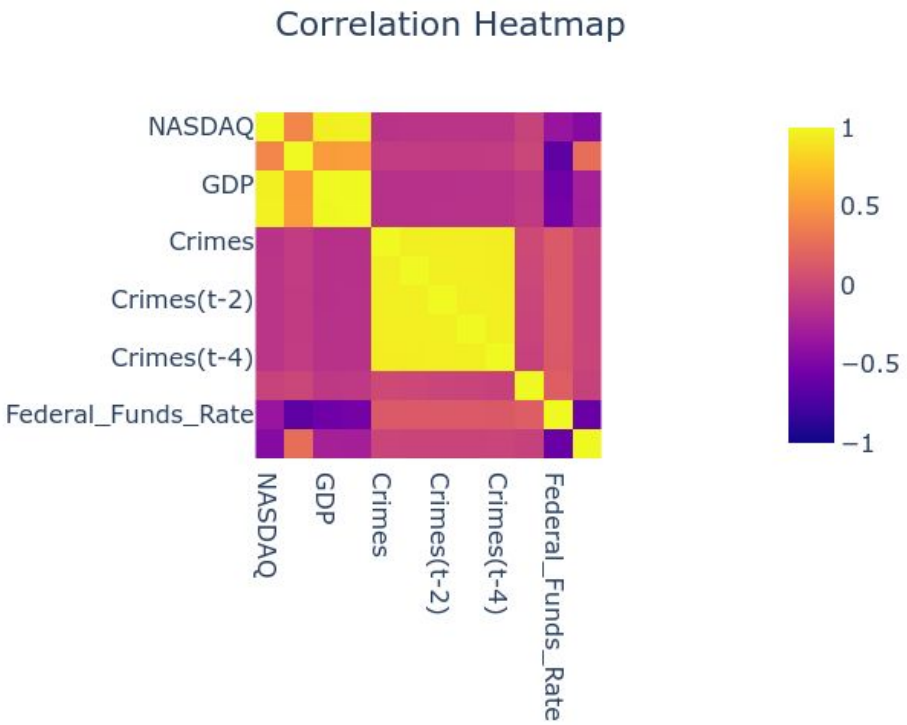
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

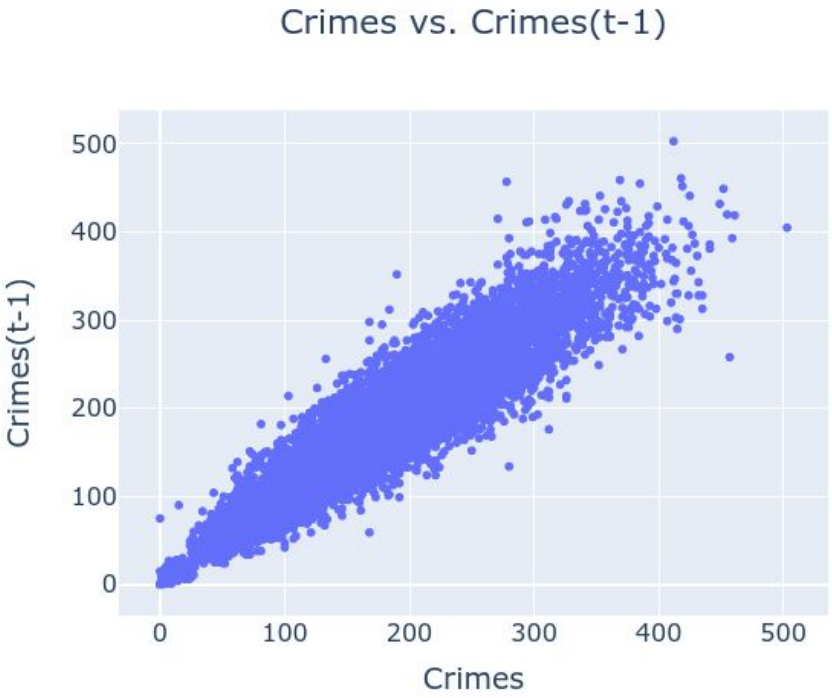
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration

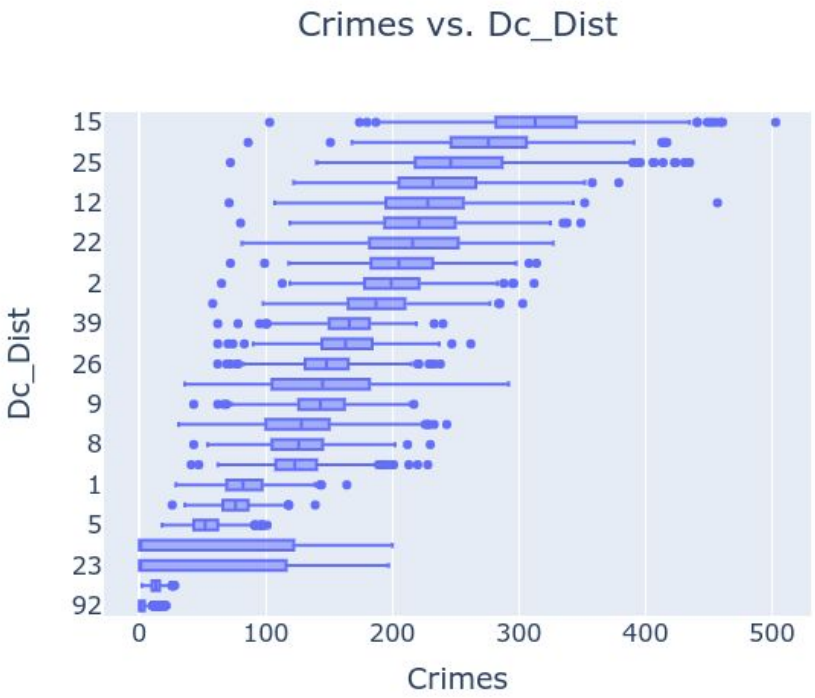




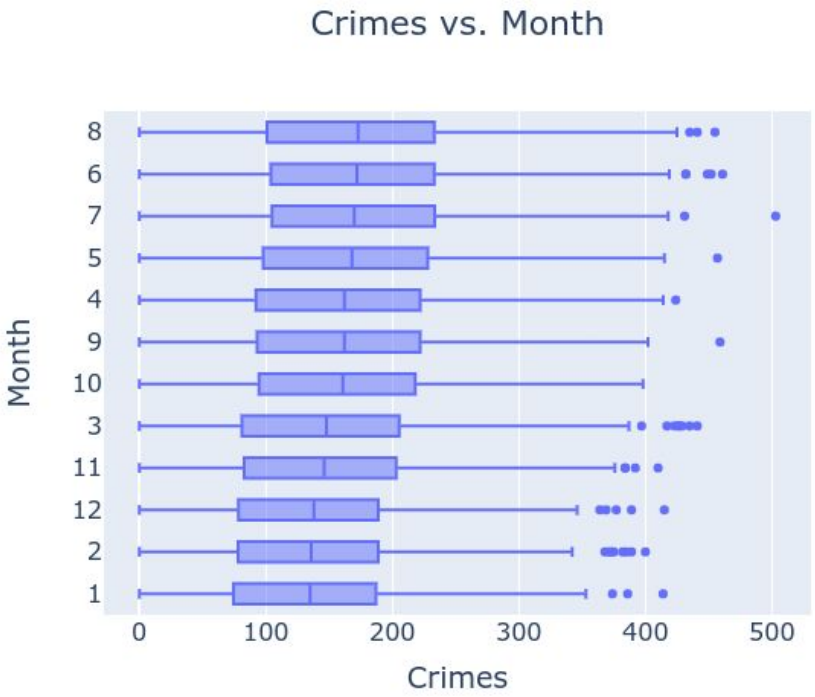
There’s two hot spots for relationships in the data. The top-left yellow region shows strong relationships between the economic indicators. The centered yellow region shows strong relationships between crime and previous weeks of crime. There’s also a strong relationship in the bottom-left where it is dark purple for the federal funds rate and the other economic indicators.



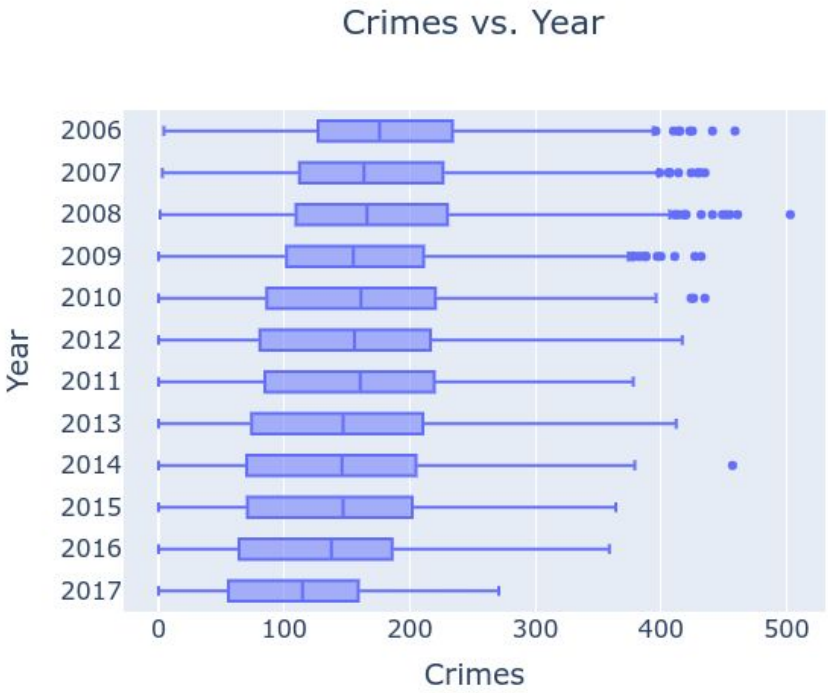
We can see that the previous week of crime is similar to the current week of crime. When the previous week is high or low the current week follows the same trend.



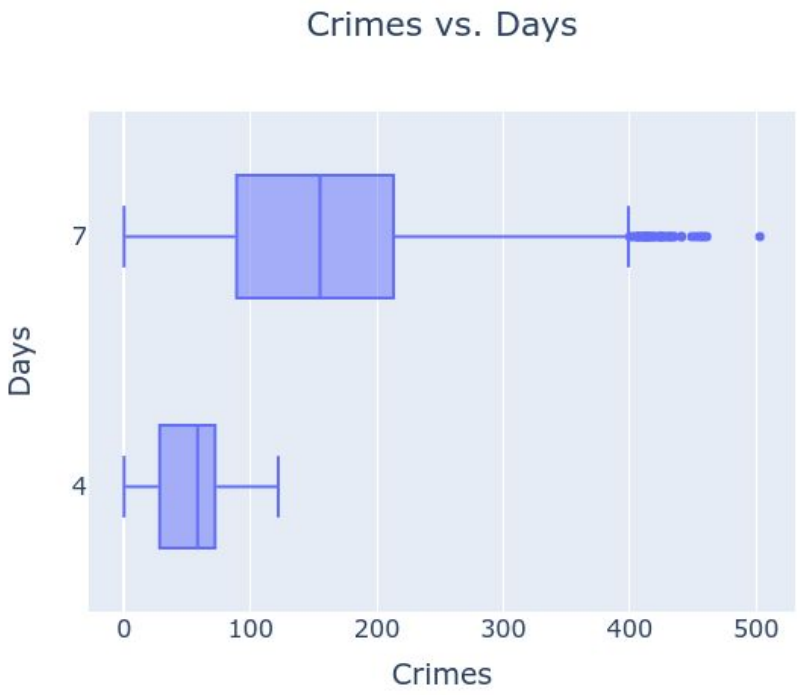
We can see that crime is higher in some districts and lower in others.



We can see that crime is slightly higher in the summer than the winter. In August, the median crime rate is 173 arrests per week. In January, the median crime rate is 135 arrests per week. That's a 22% decrease.

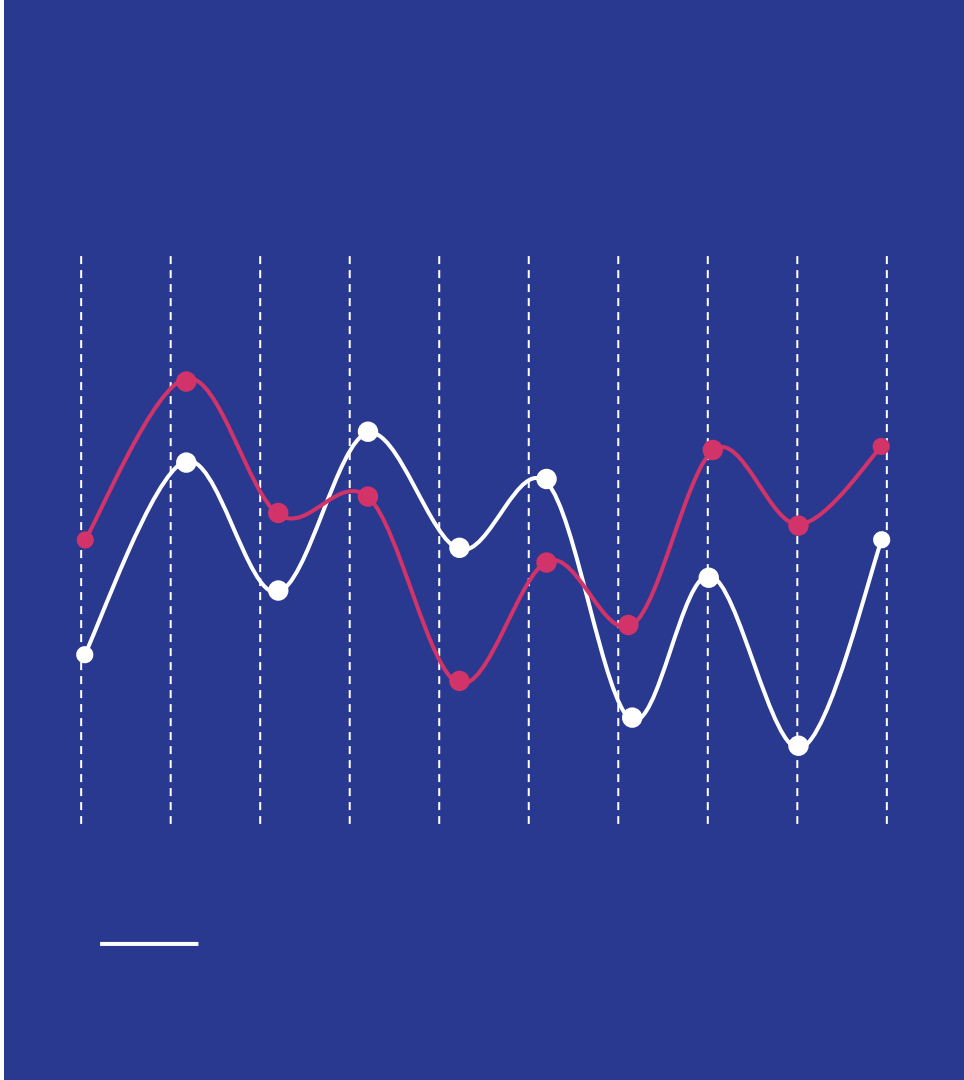


We can see that crime is slowly decreasing over time. In 2006, the median crime rate is 176 arrests per week. Ten years later in 2016, the median crime rate is 137.5 arrests per week. That's a 22% decrease.



We can see that some weeks only have 4 days in the week which results in less crime.

Modeling



Model Parameters

Linear Regression

Library: scikit-learn
Length Of Path: 1e-9
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 500

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Linear Regression

R2: 0.94
RMSE: 20.7
In Control: 97.39%

Model Indicators:
1. Crimes(t-1)
2. Crimes(t-2)
3. Crimes(t-3)
4. Crimes(t-4)
5. Days_4

XGBoost

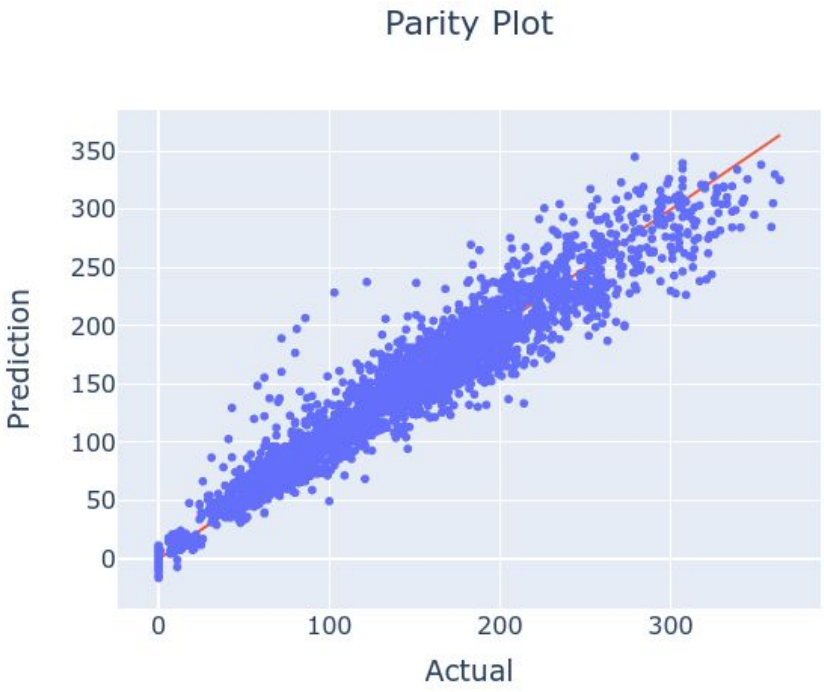
R2: 0.94
RMSE: 21.4
In Control: 97.25%

Model Indicators:
1. Crimes(t-1)
2. Crimes(t-2)
3. Crimes(t-3)
4. Crimes(t-4)
5. Days_4

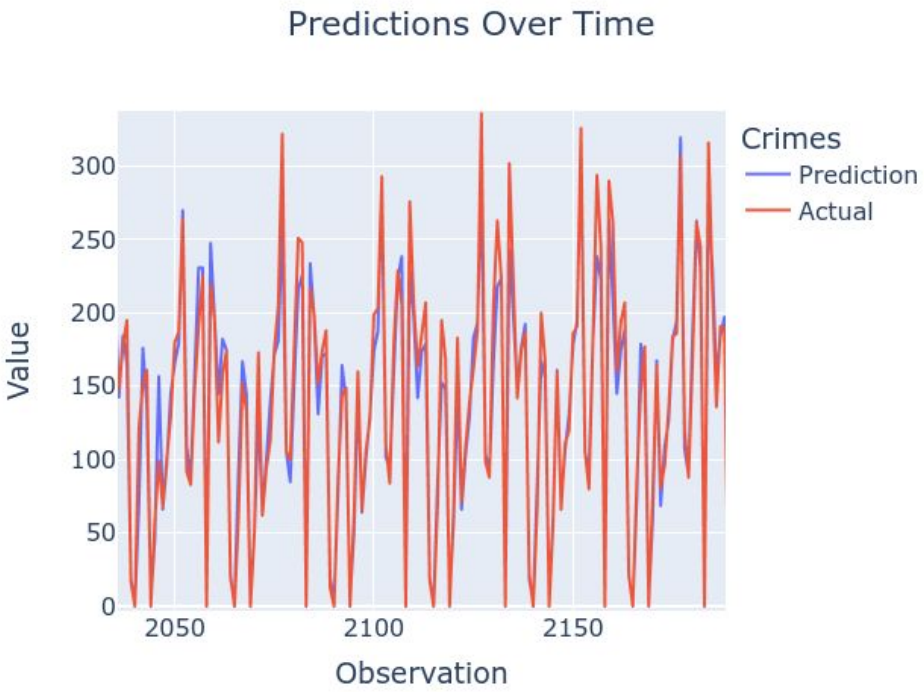
Neural Network

R2: 0.87
RMSE: 30.5
In Control: 97.43%

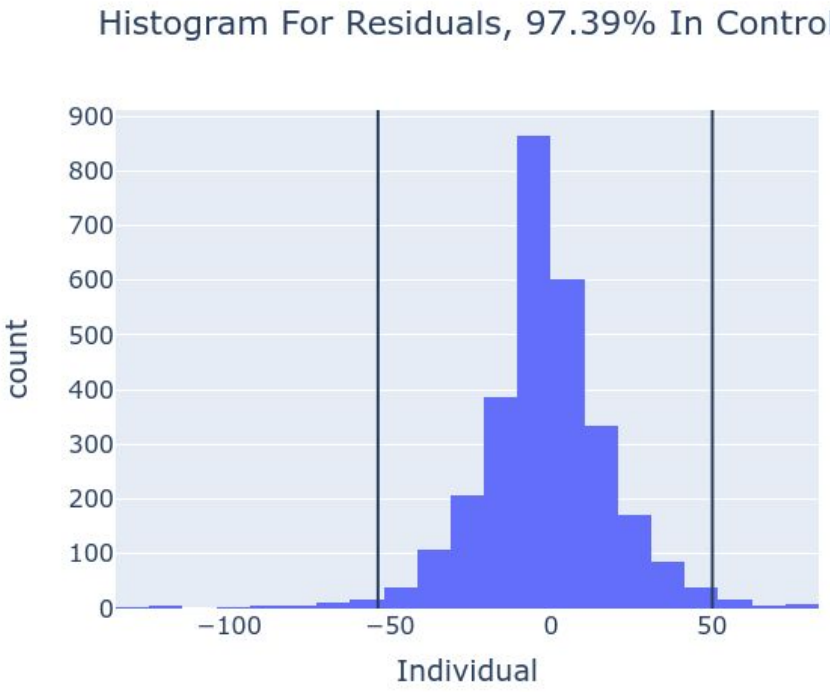
Model Indicators:
1. GDI
2. Federal_Funds_Rate
3. GDP
4. PPI
5. Unemployment



These predictions come from the linear regression model. These predictions are done on 20% of the data that the model did not see during training. The model trained on data from 2006, week 4 to 2015, week 1. These predictions are done on data from 2015, week 1 to 2017, week 12. The predictions are centered on the red line (perfect predictions). There is a tendency to over-predict the crime rate.

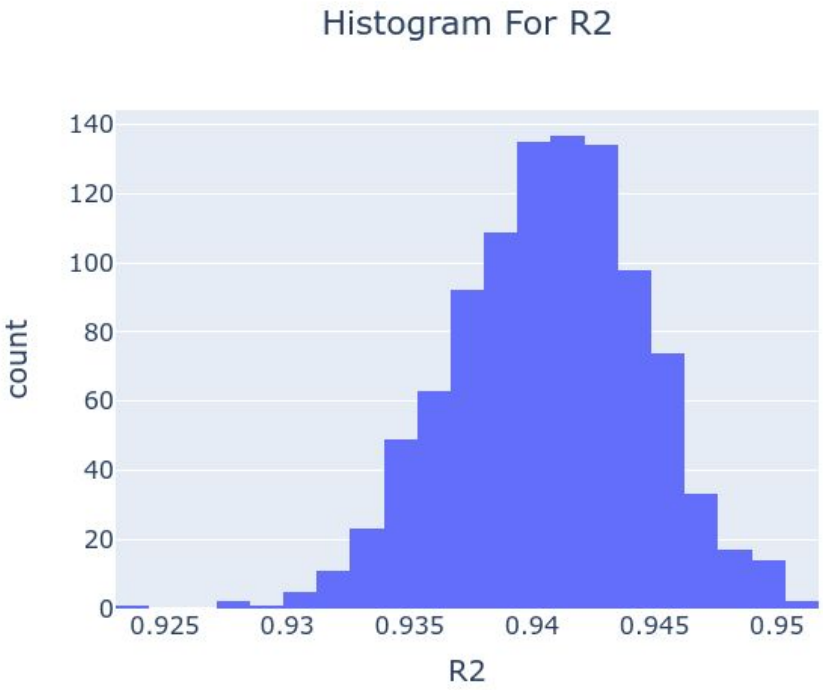


Here's a snapshot of the predictions over time. We can see the the blue predictions follow the actual values well. There is some extreme values where the red line drops down or jumps up well past the predictions. We can see there is some seasonality in the data as well.



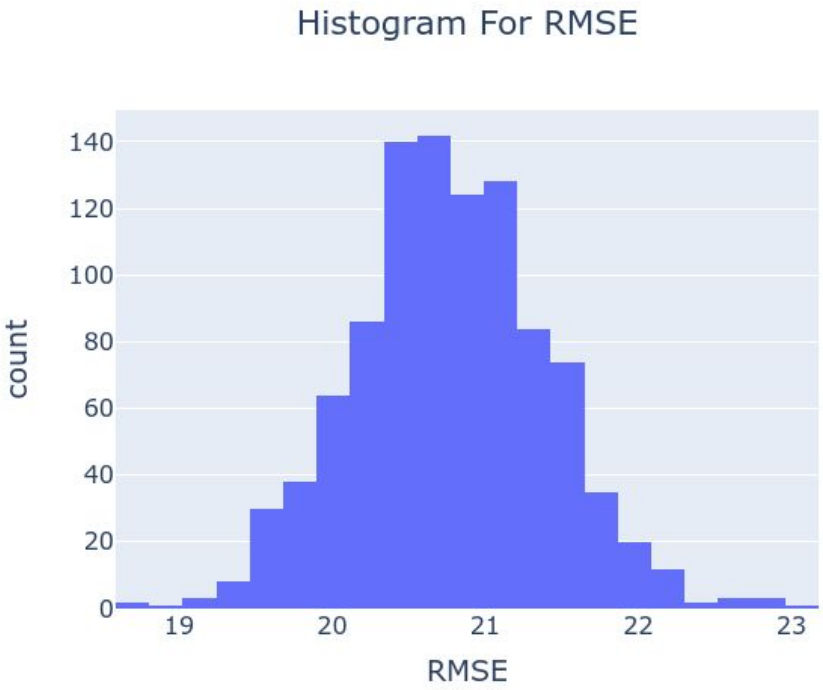
The residuals are prediction error = actual - predicted.

The residuals have a tight bell shape, which is good, and they are centered on zero. Control limits were computed on the residuals and we can see that the prediction error is mostly under control. We can see a skew to the left, which shows the model has a tendency to over-predict the crime rate.



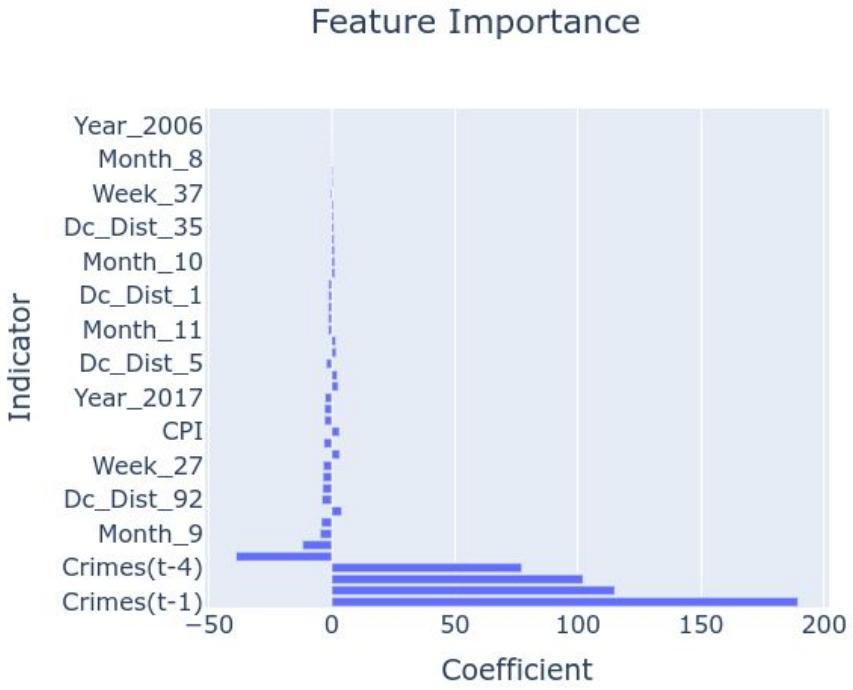
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then R2 was computed on each sample to get a distribution.

R2 has a tight range between 0.92 and 0.95, which is good. R2 has a bell shape, which is good, and a slight skew to the left.

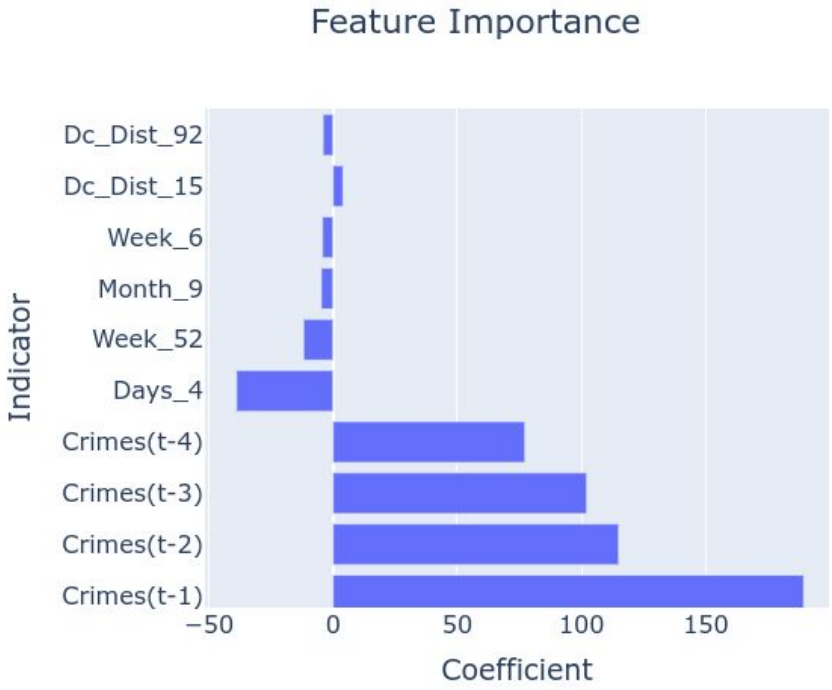


The prediction error was resampled as previously mentioned to get a distribution for RMSE.

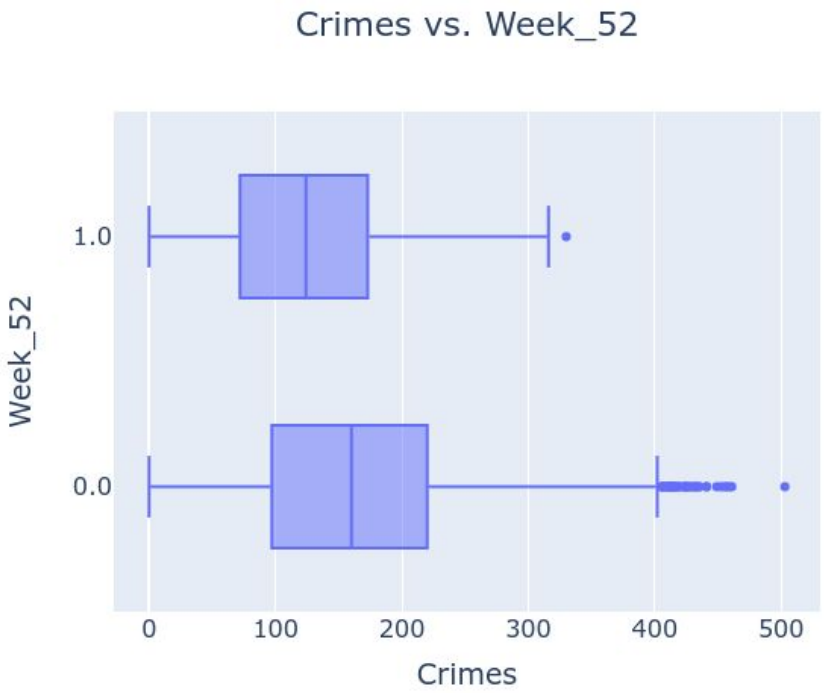
On average, the predictions are off by 19 to 23 crimes per week, which is a tight range.



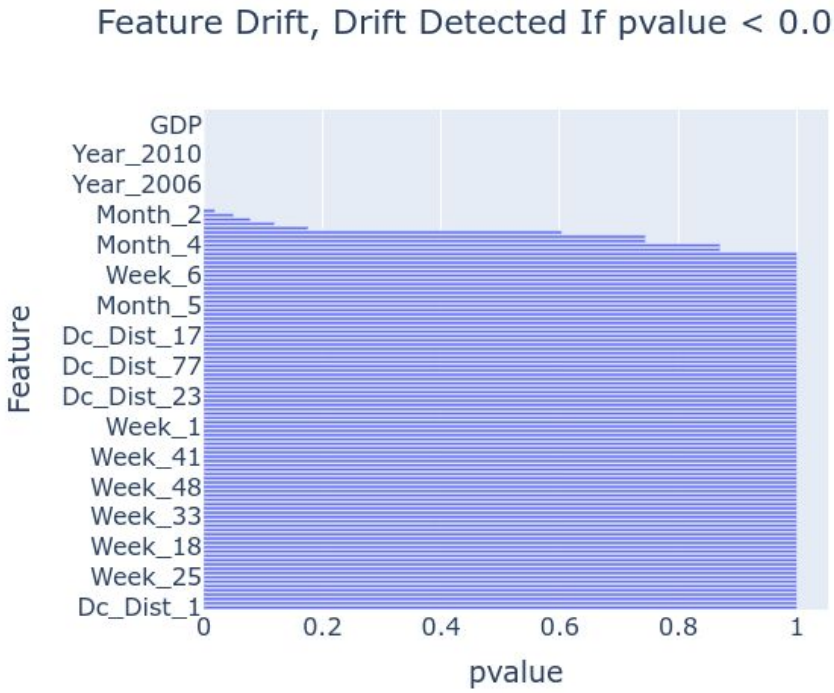
We can see that the model leverages many features to make predictions.



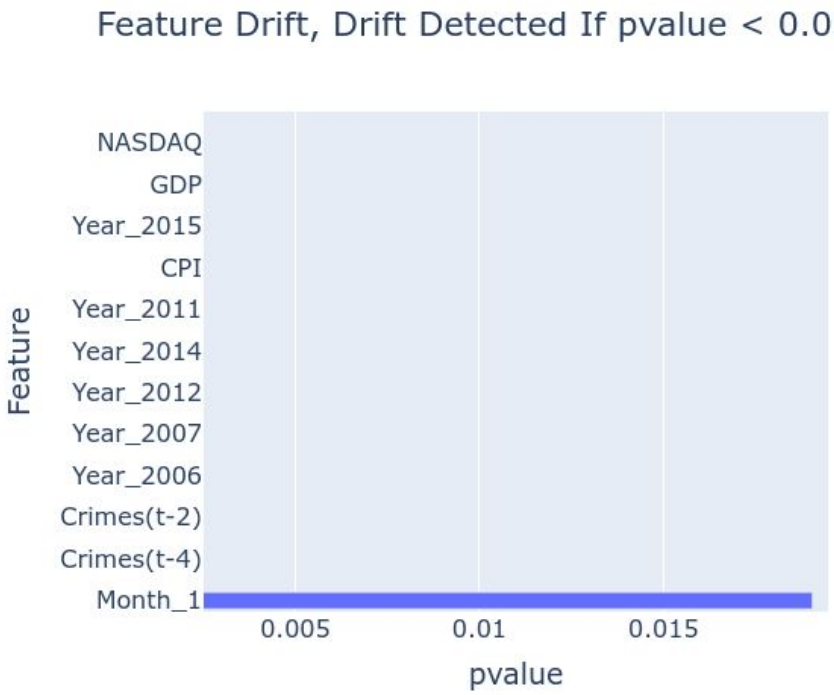
The top 10 important features are shown to the left. There is a large difference in importance between these features with the previous weeks of crime and number of days in the week contributing most to the predictions.



Week 52 is another indicator showing less crime than the other weeks. This is because week 52 is at the end of the year when the week is shorter.



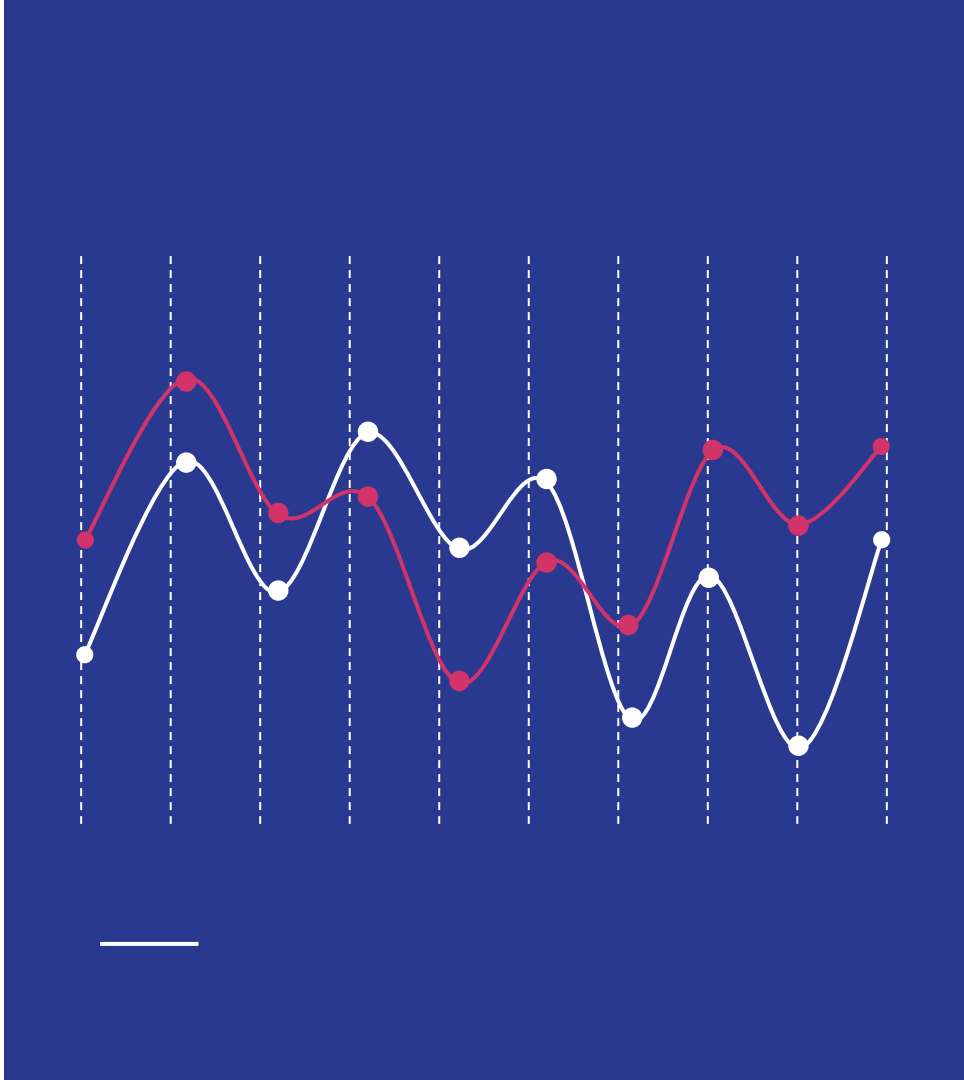
A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. Most of the columns do not experience a drift, which is good; but some of them do.

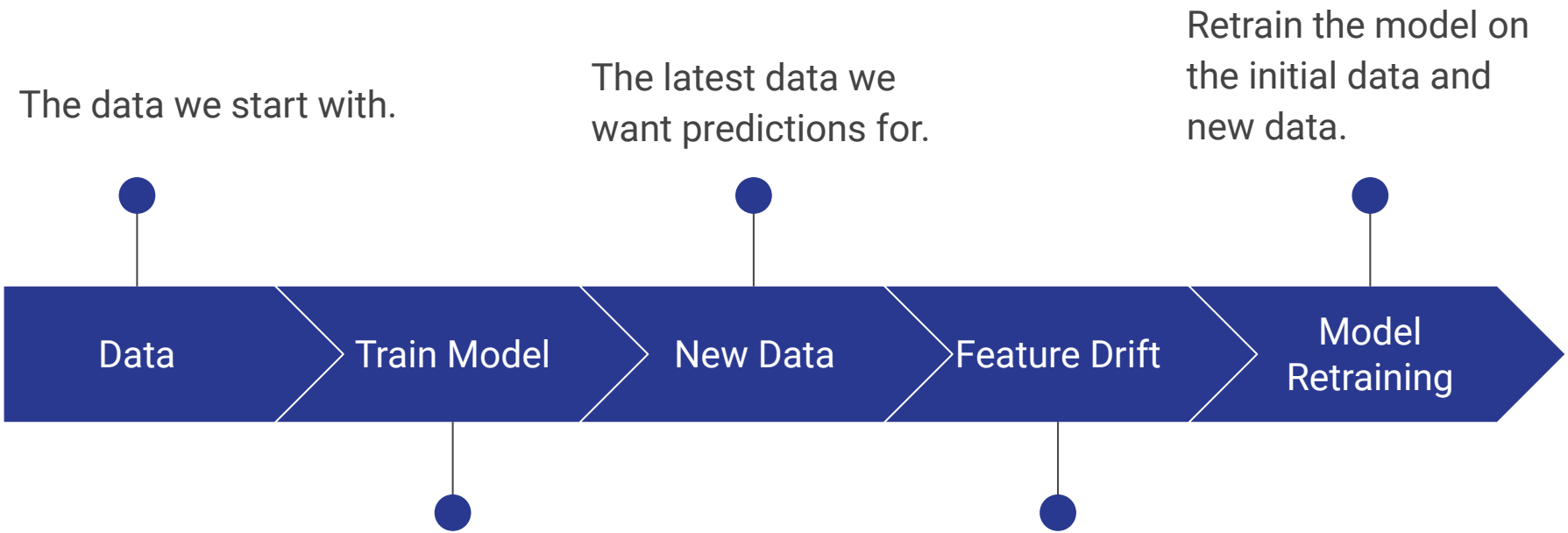


We can see that the crime rate, the years, and the economic data are experiencing a drift. This is because crime is decreasing, the years are different between the training and testing data, and most of the economic data has a trend to increase.

The linear regression model was retrained to include the testing data.

Deployment




The data we start with.
The latest data we want predictions for.
Retrain the model on the initial data and new data.
Data
Train Model
New Data
Feature Drift
Model Retraining
Data wrangling, feature engineering, model training.
See if the distribution of the new data is significantly different than the initial data.

Thank You