Forecasting Crime based on historical service call data for the city of Tempe

Satya Pranay Manas Nunna
Computer Science
Arizona State University
Tempe AZ USA
snunna7@asu.edu

Teja Naidu Koppineni
Computer Science
Arizona State University
Tempe AZ USA
tkoppine@asu.edu

Sai Neeraj Bobba
Data Science Analytics Engineering
Arizona State University
Tempe AZ USA
sbobba7l@asu.edu

Aishika Rathnala
Computer Science
Arizona State University
Tempe AZ USA
arathnal@asu.edu

Sai Harshith Reddy Muthani
Computer Science
Arizona State University
Tempe AZ USA
smuthani@asu.edu

ABSTRACT

In this urban governance initiative focused on enhancing public safety in Tempe, we employ a data-driven approach using historical service call data spanning three years. Our objective is to develop predictive models, including LSTM, Gradient Boosting Regressor, Feed Forward Neural Networks, and Random Forest to forecast crime occurrences. By examining the intricate relationship between past service call data and crime rates, we aim to provide valuable insights for law enforcement to optimize resource allocation, improve crime prevention strategies, and enhance overall public safety.

Utilizing a dataset from the city of Tempe with real-time updates ensures continuous availability of fresh samples for model evaluation. The project involves thorough train-test splits on existing data to assess the accuracy of predictive models against real-time updates, allowing for a comprehensive evaluation of practical utility in real-world scenarios.

The primary problem addressed by this project is the need for efficient deployment of police forces based on a detailed analysis and understanding of crime patterns in Tempe. Crime analysis and forecasting are crucial aspects of urban governance, and this project contributes to the collective effort of cities aiming to identify crime hotspots. By providing a solution to the crime forecasting problem through the analysis of Tempe's calls for service data, we strive to assist crime analysts in making cities safer.

It's worth noting that the dataset includes various service calls beyond crime data, such as wellness checks. However, the ultimate goal is to identify locations generating most service calls for optimal personnel distribution. The project aims to enable law enforcement to adopt a proactive approach, potentially improving response times and promoting a paradigm shift towards data-driven policing. By leveraging advanced analytical techniques, this analysis seeks to uncover pattern trends, offering a valuable tool for addressing urban crime challenges and contributing to the evolution of law enforcement strategies.

INTRODUCTION

Addressing criminal activity is a critical concern in urban areas, as it significantly impacts citizens' peace and freedom of movement. Crimes like robberies, burglaries, and murders create a sense of unease within communities. Law enforcement faces the ongoing challenge of identifying areas prone to future criminal incidents. If they could preemptively acquire this information, they could allocate resources strategically, enhancing safety and community security.

This project seeks to tackle this issue through a data mining initiative, employing predictive modeling methods to forecast crime incidents. The Tempe Call for Service dataset is utilized, containing crucial information such as event dates, grid IDs, geographical coordinates, call types, postal codes, zones, and priority levels. The first phase involves organizing the data by extracting relevant information based on grid ID, date and time, and the

number of reported offenses in each location, forming the foundation for the predictive model.

The predictive system is trained to anticipate future criminal activities using this refined dataset. Advanced algorithms analyze past data to uncover trends and patterns that could be utilized to estimate the location and time of crimes. This technology aims to assist law enforcement agencies proactively in identifying potential hotspots for criminal activity. A heat map will be developed to pinpoint areas where criminal activity is more likely to occur. This map will use distinct markers to indicate hotspot areas with varying frequencies and severity based on the number of criminal occurrences in specific locations.

The dataset from the city of Tempe is used because it is regularly updated, ensuring a consistent supply of fresh data for ongoing model evaluations.

RELATED WORK

Over the past few years, various research studies have made significant strides in the field of crime forecasting and analysis. For instance, Yu and colleagues [1] conducted an extensive investigation into data mining techniques for predicting crime, including decision trees, clustering, and neural networks, evaluating their predictive accuracy. Borowik and team [2] delved into time series analysis methods, such as ARIMA, STL, ETS, Prophet, and LSTM, addressing the challenges associated with forecasting crime trends over time. Sathyadevan et al. [3] proposed a distinctive system that utilized data mining methods to predict crime-prone areas and visualize them using a heat map technique. Alghamdi and Al-Dala'in [4] introduced an innovative approach involving deep neural networks for predicting crime activities, optimizing hyperparameters for spatiotemporal attributes, and dealing with imbalanced data issues. Yadav and Sheoran [5] presented a new approach to crime prediction using Auto Regression Techniques, considering spatial-temporal patterns of crime events and the correlation between crime attributes and expected crime levels. Feng et al. [6] employed big data analytics and data mining techniques to analyze crime data from major U.S. cities, utilizing advanced models like the Prophet model and LSTM neural networks. Jiang et al. [7] merged LSTM-based crime prediction with spatial analysis, incorporating Point of Interest (POI) data and crime record data to explore spatio-temporal dynamics. Kim et al. [8] investigated crime prediction using machine learning, refining the dataset and providing insights into the effectiveness of K-Nearest Neighbor and boosted decision tree models. Biswas et al. [9] tackled rising crime rates in Bangladesh using regression models, including linear regression, polynomial regression, and random forest regression, to predict crime trends based on nearly two decades of data. In a separate study, Kim et al. [10] integrated the Safe Route Travel app and Crime Mapping system with fuzzy-based crime prediction using the Fuzzy K-Nearest Neighbor (FKNN) algorithm, incorporating real-time crime alerts based on historical records for the city of Baltimore, Maryland, USA. Collectively, these studies enrich our understanding of crime forecasting and analysis by providing diverse methodologies, datasets, and insights that contribute to ongoing research in the field.

Most of the relevant research papers that we observed worked on obtaining a spatio-temporal analysis over a given period of time for a given location. In this project we also tried to recreate the similar analysis by creating our own spatial plotting over a specific time period for the city of tempe. We tried to add appropriate bias variables in order to understand their effect on the number of crime incidents occurring.

DATA

The dataset that has been used for the analysis calls for service data of the city of tempe for the years {2020 to 2023 [Up to Oct 22nd]}.

The columns that are available in the dataset are displayed below and the shape of the dataset is (428367, 37).

```
Index(['X', 'Y', 'OBJECTID', 'PrimaryKey', 'OccurrenceDatetime',
       'OccurrenceYear', 'OccurrenceMonth', 'OccurrenceHour', 'OccurrenceWeek',
       'OccurrenceDatePart', 'OccurrenceWeekday', 'ObfuscatedAddress',
       'FinalCaseType', 'Priority', 'Zone', 'Grid', 'HowReceived', 'UnitID1',
       'UnitID2', 'UnitID3', 'ReportFlag', 'CaseStatus', 'ClearedBy',
       'XCoordinate', 'YCoordinate', 'Disclaimer', 'PlaceName', 'CallType',
       'Latitude', 'Longitude', 'CharacterArea', 'ReportDistrict',
       'ReportBeat', 'PostalCode', 'CensusTractID', 'ParkName',
       'NeighborhoodName'],
      dtype='object') (428367, 37)
```

The examination of the dataset highlights significant connections, especially within the geographical and temporal aspects. A noteworthy finding is the moderately positive correlation between OccurrenceYear and Priority, suggesting a potential trend in how incident priorities are assigned or the nature of incidents over time. Geographical coordinates (XCoordinate, YCoordinate, Latitude, and Longitude) exhibit nearly perfect correlation, indicating they likely represent the same locations in different coordinate systems. This is further emphasized by their moderate correlation with PostalCode, underscoring the geographical relevance of postal codes.

Temporal variables like OccurrenceMonth and OccurrenceWeek demonstrate a high correlation, which is expected given their inherent time-based relationship. However, their correlation with the hour of occurrence is minimal, suggesting that the time of day incidents occur is independent of broader temporal trends. The dataset also suggests a subtle geographical variation in incident priority,

supported by a weak negative correlation between Priority and Longitude.

In summary, while the dataset provides valuable insights into linear relationships among its features, additional analysis is required to uncover underlying patterns and causal connections.

Dropping the Unnecessary Data:

Dropping the data which is not necessary. For instance, duplicate columns like ‘X’, ’Y’ (Latitude, Longitude) are dropped. Similarly columns such as 'OBJECTID', 'PrimaryKey', 'ObfuscatedAddress', 'CensusTractID', 'ParkName', 'NeighborhoodName', 'FinalCaseType', 'HowReceived', 'UnitID1', 'UnitID2', 'UnitID3', 'ReportFlag', 'CaseStatus', 'ClearedBy', 'Disclaimer', 'PlaceName', 'CharacterArea', 'ReportDistrict', 'ReportBeat’ were dropped due to the data containing uniqueIDs, labels, and other data that is not required in aggregating the data.

EDA (Exploratory Data Analysis):

The dataset that has been refined are displayed below:

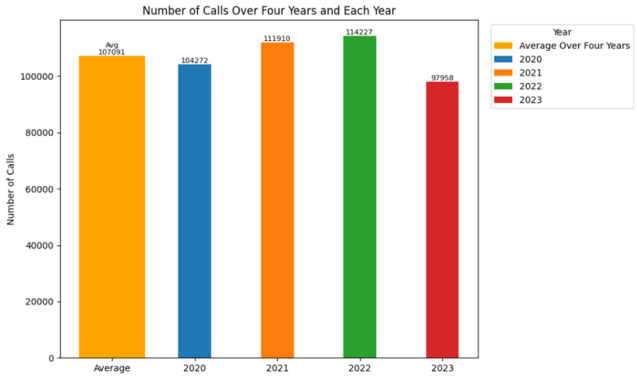
OccurrenceDatetime	OccurrenceYear	OccurrenceMonth	OccurrenceHour	OccurrenceWeek	OccurrenceDatePart	OccurrenceWeekday	Priority	Zone	Grid	XCoordinate	YCoordinate	CallType	Latitude	Longitude	PostalCode
2023-10-23 0:35:24	2023	10	0	43	23	Monday	2	19	1304	660674	875646	Trespassing Call	33.407069	-111.9462447	85282
2023-10-23 0:55:12	2023	10	0	43	23	Monday	2	DT	709	660969	881225	Other Officer Initiated Call	33.422423	-111.9397152	85281
2023-10-23 0:17:12	2023	10	0	43	23	Monday	3	21	2503	666908	863982	Loud Music/Noise/Party Call	33.37250302	-111.9467768	85283
2023-10-23 0:49:14	2023	10	0	43	23	Monday	2	13	816	660486	875754	Other Officer Initiated Call	33.40738582	-111.947844	85282
2023-10-23 1:50:01	2023	10	1	43	23	Monday	2	19	1901	660715	870400	Drunk Disturbing Call	33.39267226	-111.9455355	85282
2023-10-23 1:03:01	2023	10	1	43	23	Monday	4	14	1010	701227	877920	Illegal Parking Call	33.41334252	-111.9126455	85281
2023-10-23 2:02:36	2023	10	2	43	23	Monday	2	DT	709	660883	882298	Subject Disturbing/Harassing Call	33.42537289	-111.9396638	85281
2023-10-23 2:15:36	2023	10	2	43	23	Monday	2	15	1111	702305	875799	Other Officer Initiated Call	33.40751299	-111.9091132	85282
2023-10-23 2:49:48	2023	10	2	43	23	Monday	2	24	3107	696729	849621	Cancelled Call	33.33566456	-111.9273769	85284

The dataset focuses on individual instances of criminal activities, providing detailed information about each specific incident. This includes the location where the crime occurred, as well as the date and week in which it took place. The dataset's granularity is at the level of singular crimes, offering a comprehensive account of each separate occurrence. There is a grid parameter which doesn't have any corresponding grid mapping file so we decided to create our own grids according to the tempe city coordinates. Analyzed the dataset to get to know how the distribution of the data over multiple cohorts took place.

Finding Trends:

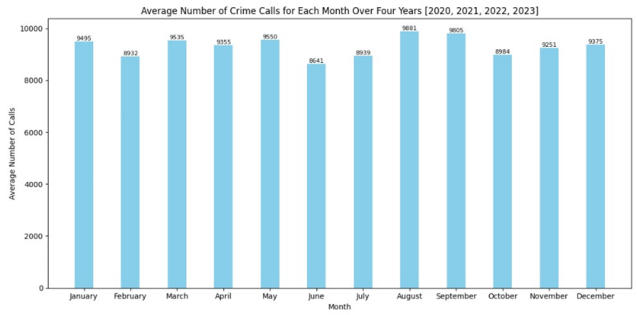
Step 1: Year vs Number of Incidents : [2020 - 2023 {Till Oct 22nd}]

As per the below graph, there are an average of 107091 crime calls over the four years, but there is no observable increasing or decreasing trend.



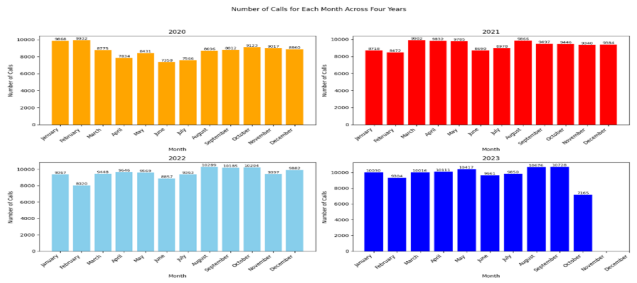
Step 2: Month vs Average Number of Incidents:

As per the below graph, there is no noticeable increasing or decreasing trend over the period of four years.



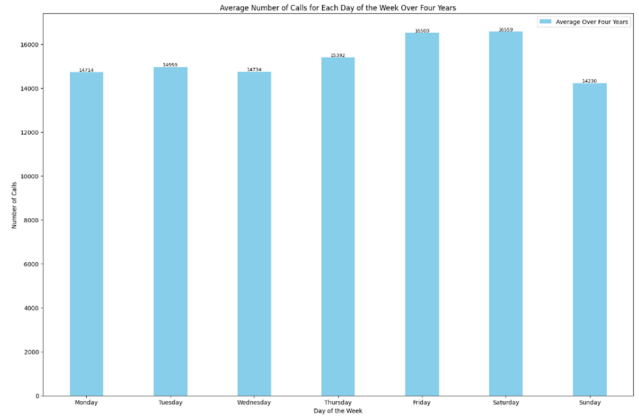
Step 3: Month vs Average Number of Incidents over the four years

This graph as well shows no observable trend



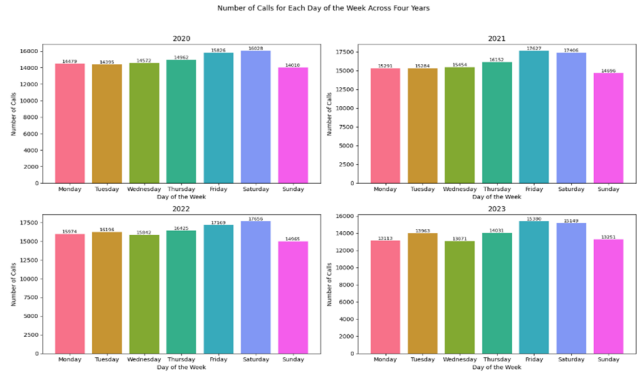
Step 4: Weekday vs Average number of incidents

From the below graph we can observe that there are a huge number of incidents happening during Friday and Saturday.



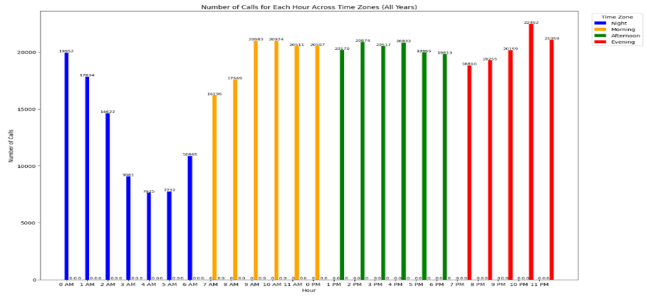
Step 5: Weekday vs Number of Incidents over 4 years:

From the below graph we can observe the same thing as above without any difference in trend like Friday and Saturday are the days having huge numbers of crime calls.



Step 6: Time vs Number of Incidents:

From the below graph we can see that the calls tend to go down during the early hours then have a constant occurrence rate with a spike at 10 - 11 pm. (Including this factor would cause having sparse data).



Final Observations:

We decided to move forward and aggregate the data at day, month and yearly levels and chose to add a spatial component in terms of the grid. There is already a grid value that is available in the dataset which does not have grid mapping so we decided to make our own grids. We have seen an observable trend in the weekdays graph so we considered that as a potential bias variable. The Latitudes and Longitudes are going to be used to find the grid where each call belongs to.

METHODS

After completing the literature survey we understood that the best possible approach to effectively predict crime incidents would be to perform a spatio-temporal analysis. We already have the temporal factors in our available dataset in order to create the spatial factors we will be dividing the geographical area into grids.

Creating grids:

We intend to establish a mapping system over the Tempe area using boundary data specific to the region. Our decision to implement a 7x7 grid is based on the consideration that this resolution strikes a balance between capturing detailed geospatial trends and avoiding sparsity issues with data. Increasing the resolution could lead to insufficient data per grid, while decreasing it might result in overly generalized patterns.

In this grid system, each cell is uniquely identified by a grid number, and it corresponds to a specific polygon defined by its X-min, X-max, Y-min, and Y-max values. This approach ensures that our mapping is finely tuned to the intricacies of the Tempe landscape, allowing us to effectively leverage geospatial trends in our model training process.

Mapping Grids to Data and Aggregation

We've established a mapping system that links the grids to specific call locations using their respective x and y coordinates. To enhance our analysis, we aggregated the initially individual crime data into a consolidated format, grouping it by grid number, occurrence year, occurrence month, and occurrence date. As part of this aggregation, we transformed the data from a single crime level to a count of crimes within each grid, during each year, month, and date of occurrence.

To retain the spatial aspect of the data, we introduced x-centroid and y-centroid values. These represent the centroids of each grid, providing a central reference point. Our objective is to leverage the available parameters, including the spatial information, to predict the number of occurrences within each grid. This predictive modeling approach aims to enhance our understanding of crime patterns and trends based on the specified parameters.

The parameters that are available for prediction:

	GridNumber	OccurrenceYear	OccurrenceMonth	OccurrenceDatePart	NumberOfIncidents	GridID	X_centroid	Y_centroid
0	1.0	2020	1	1	1	1	-111.971255	33.330417
1	1.0	2020	1	2	1	1	-111.971255	33.330417
2	1.0	2020	1	3	6	1	-111.971255	33.330417
3	1.0	2020	1	4	5	1	-111.971255	33.330417
4	1.0	2020	1	5	1	1	-111.971255	33.330417

EXPERIMENTS

Hypothesis: Our aim is to predict the number of crimes in each grid by analyzing historical data, with a specific focus on temporal factors such as the year, month, and day.

Model Selection: We plan to explore various predictive models to achieve our goal. The potential machine learning models that are considered are Random Forest, Gradient Boosting Regressor, LSTM (Long Short-Term Memory), and Feed Forward Neural Network.

Model Development:

In the initial phase, we strategically employed a combination of cross-validation (CV) and an 80:20 split to partition our dataset. Recognizing the importance of considering the temporal aspect, we aimed to enhance the robustness of our data division for training and testing purposes. This dual approach allowed us to leverage the benefits of both cross-validation techniques and a straightforward 80:20 split, contributing to a more comprehensive evaluation of our model's performance.

Model Training: Following that, we trained our selected models using the training set. This step involved feeding the historical data into the models, allowing them to learn patterns and relationships.

Hyperparameter Tuning: To enhance the performance of our models, we engaged in hyperparameter tuning. This process entailed optimizing the model parameters through a grid search, systematically exploring various combinations to identify the most effective configuration.

By following this comprehensive approach, we aimed to develop robust predictive models that could accurately forecast the number of crimes in each grid based on historical data and temporal factors.

Model evaluation: Assessing the effectiveness of regression models involves employing metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R2 score. To ensure the robustness of the models, a 5-fold cross-validation technique is implemented.

Validation: For the validation phase, the model is tested on unseen data spanning from October 23rd, 2023, to November 12th, 2023. This step aims to verify if the model's predictions align closely with the actual observed data.

Model Interpretation: In terms of model interpretation, an analysis is conducted to comprehend the significance of various features such as grid, year, month, and day in predicting crime rates. This process involves exploring how changes in these features impact the model's predictive capabilities.

In our initial hypothesis we intended to find the number of incidents that could occur in an area over a period of time after performing the above experiment we were able to analyze and understand the number of crime incidents over a specific period of time, thus experiment were able to works towards our goal of predicting the crime incidents in the city of tempe.

Additionally, a bias variable, namely "weekday," is introduced to the model. The objective is to evaluate how the inclusion of this variable affects the model's performance. This step helps in understanding whether the model adapts well to variations influenced by the day of the week, providing insights into the role of weekdays in predicting crime rates.

RESULTS

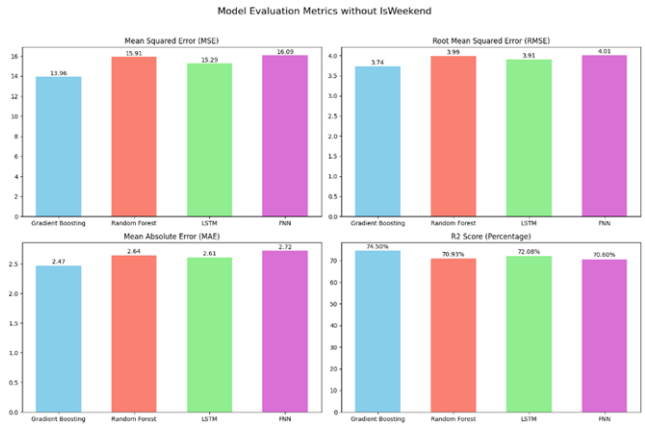
In our experimental setup, we conducted a comprehensive analysis involving the training of four distinct models: Random Forest, Gradient Boosting Machines (GB), Long Short-Term Memory (LSTM), and Feed Forward Neural Network. To assess the models' effectiveness, we initially employed an 80:20 train-test split on the data.

Subsequently, to check the robustness of these models, we implemented a 5-fold cross-validation technique. This involved dividing the dataset into five subsets, training the models on four of these subsets, and validating their performance on the remaining subset. This process was repeated five times, with each subset serving as the validation set exactly once. The evaluation of model performance was based on key metrics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared (R2) score.

In a specific analysis excluding weekdays, we observed that the Gradient Boosting (GB) Regressor emerged as the top-performing model according to the graphical representation of results. In an attempt to enhance the R2 score, a bias variable was introduced, and the outcomes of

this modification are presented in detail on the subsequent slide.

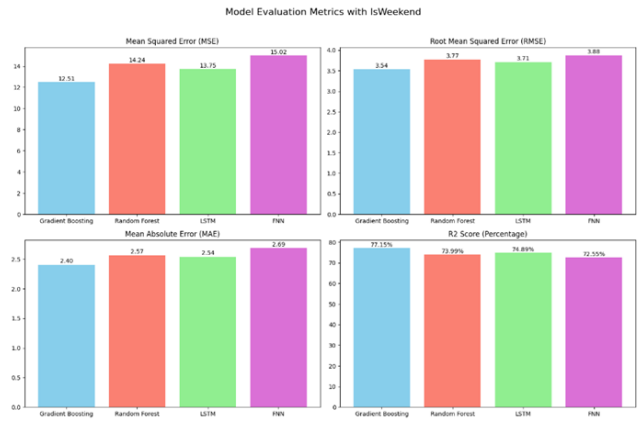
Model evaluation metrics without IsWeekend:



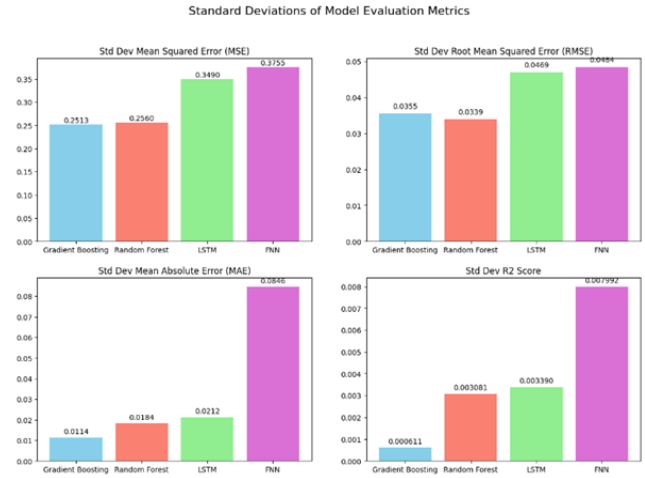
Standard Deviation:

The analysis reveals that the GB Regressor outperforms other models across all categories. Introducing weekdays as a bias variable has proven to be a significant factor, contributing to a notable 3% improvement in the R2 score specifically for the GB Regressor. This indicates that considering the day of the week as a variable has positively impacted the model's predictive accuracy, showcasing the importance of incorporating this additional information for better performance.

Model evaluation metrics with IsWeekend:



Standard Deviation:



Spatial Representation:

The R2 score for the Gradient Boosting Regressor model during the period from October 23rd to November 12th, 2023, is 71.17%. This statistical metric is indicative of the model's ability to explain the variance in the observed data. A higher R2 score suggests a better fit of the model to the actual values. In this case, the R2 score of 71.17% implies that approximately 71.17% of the variability in the data can be explained by the model.

Gradient Boosting Regressor for Test Dataset:

After performing testing using gradient boosting regressor model for the test dataset we got the below results:

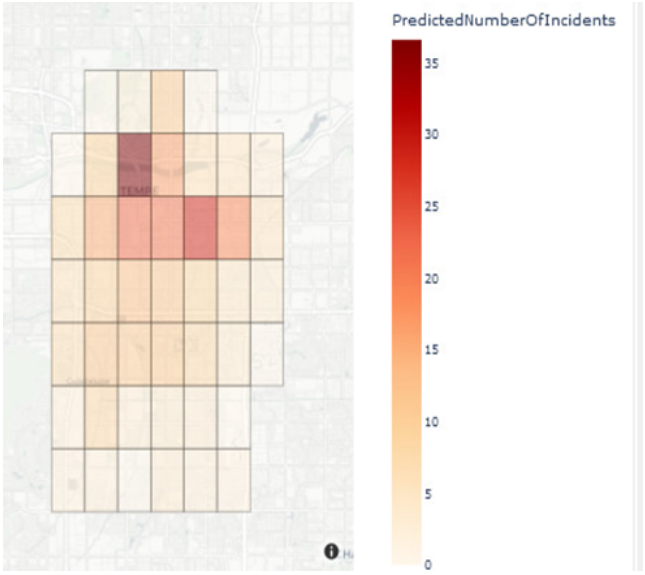
MSE - 20.58, RMSE - 4.53, MA - 2.81E, R2 Score - 0.71

```
(20.587886824914992, 4.537387665266766, 2.815545620648096, 0.7117238806125719)
```

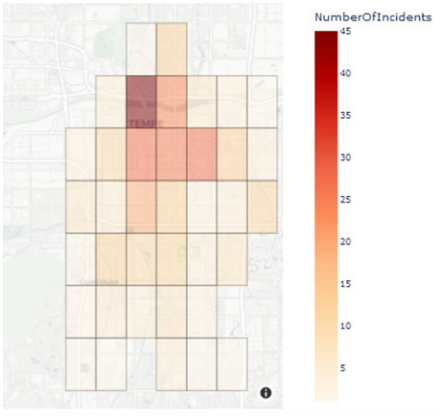
To further illustrate the model's performance, a visual representation is provided in the form of a sample graph comparison. This graph likely depicts the predicted values generated by the Gradient Boosting Regressor alongside the actual observed values for the specified time period. The comparison allows for a qualitative assessment of how well the model aligns with the real data points.

In summary, the R2 score and the accompanying graph comparison collectively provide insights into the accuracy and effectiveness of the Gradient Boosting Regressor model in capturing the patterns and trends present in the given dataset during the mentioned time frame.

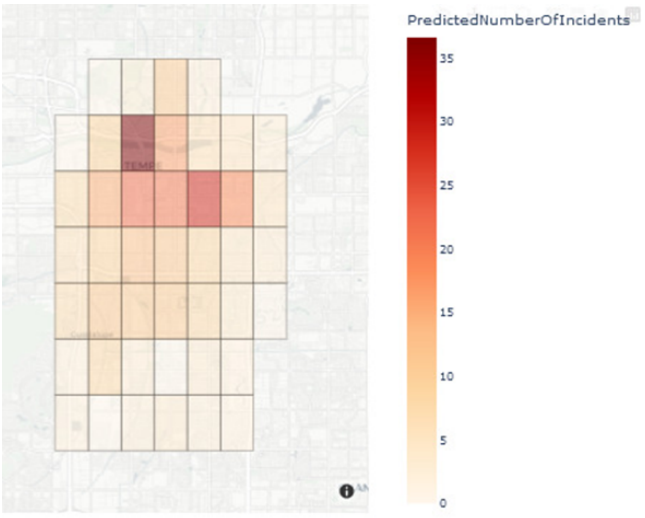
For Nov 11th 2023 (Predicted)



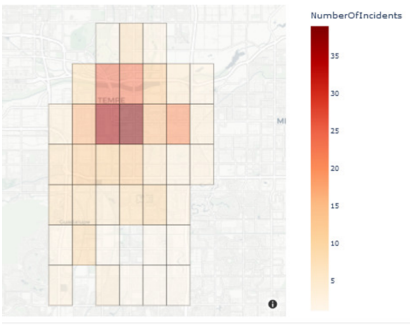
Nov 11th 2023 (Actual)



For Nov 10th 2023 (Predicted)



Nov 10th 2023 (Actual)



The model has demonstrated a commendable R2 score of 71.17%, signifying its capability to account for 71.17% of the variability in the target variable. This underscores a strong predictive accuracy. Additionally, upon scrutinizing the spatial representation, it becomes evident that while the actual number of incidents may vary, the heat map consistently highlights specific regions as potential high-risk areas.

Moreover, it was observed that the inclusion of the bias variable, specifically the one related to weekdays (IsWeekday), contributed to a noteworthy enhancement of 3% in the R2 score of the most effective model. This implies that considering the bias variable in the model significantly improved its performance, emphasizing the relevance and impact of such factors on the overall predictive accuracy.

Interesting Findings: (working only with temporal data)

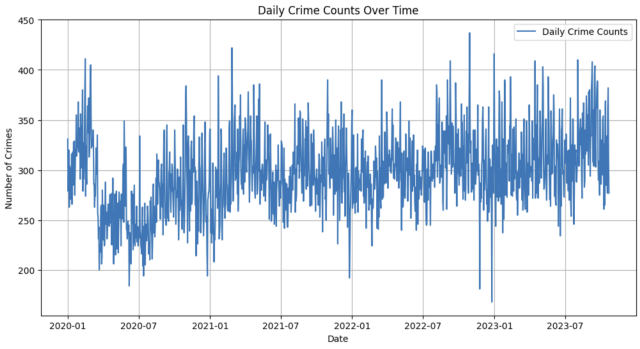
In our pursuit for unraveling patterns within crime data, we were curious about the trends on crime count on each day over a period of time. To explore this further we wanted to

predict the number of crime occurrences that could happen each day over a period of time. We had thought this could potentially further help the forces to prepare earlier if there was an increase on predicted crime rates. While this wasn't our main topic of research it did produce some results so we are going to discuss them here.

The dataset we used was the tempe call for service data. This dataset had raw information on criminal activities and has gone through a proper preprocessing phase. In the preprocessing journey we had removed irrelevant attributes like spatial coordinates, object identifiers, and redundant information. To add to this cleaning phase, any entries with missing values were removed in order to ensure the integrity of the dataset.

After this the temporal aspect of the dataset was emphasized. The “OccurrenceDatetime” in the dataset was actually in string format in the original dataset, so we converted it to the datetime format to help with the temporal analysis. By doing this transformation we were able to derive more information like day, month, year, and weekday.

Now to help the temporal analysis at a daily granularity, we had aggregated the data to get the daily crime count. This aggregation involved counting the number of crime occurrences each day leading to the creation of “dailt_crime_count” dataset. consisting of two columns ‘Date’ and ‘crime count’.

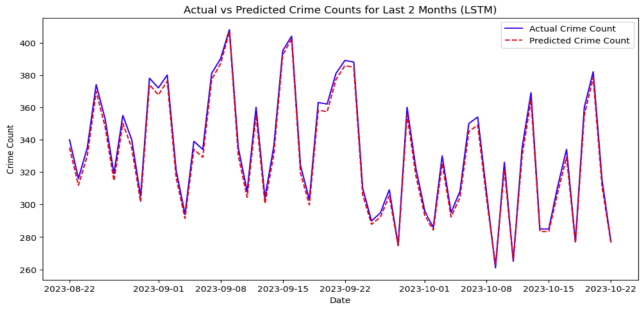


This input was used in multiple models like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), XGBoost, AutoRegressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA).

To start the analysis we first decided to use the Long Short-Term Memory (LSTM) model. We had trained the model on the dataset which has the last 60 days data removed and eventually tested it on how well it does on the last 60 days data. All our models were evaluated with Mean Squared Error (MSE), mean absolute error (MAE) and Root Mean Squared Error (RMSE). LSTM had MSE: 14.87, MAE: 3.63, RMSE: 3.86.

LSTM

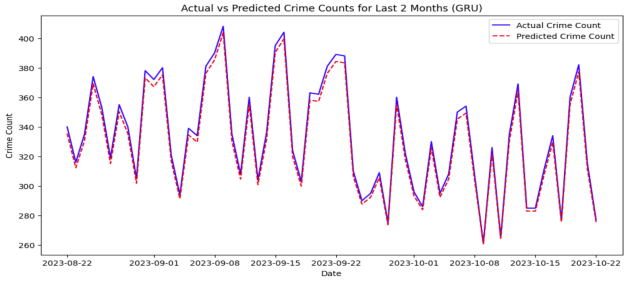
MSE	MAE	RMSE
14.87	3.63	3.86



In parallel Gated Recurrent Unit (GRU) model was introduced. This model had evaluation values: MSE: 15.74 MAE: 3.79 RMSE: 3.97.

GRU

MSE	MAE	RMSE
15.74	3.79	3.97



After this we explored using XG-boost, however the models performance wasn't great compared to the others and we didn't have enough time to to improve its performance. Its performance was like this: MSE: 998.17 MAE: 26.97 RMSE: 31.59.

XG-Boost

MSE	MAE	RMSE
998.17	26.97	31.59

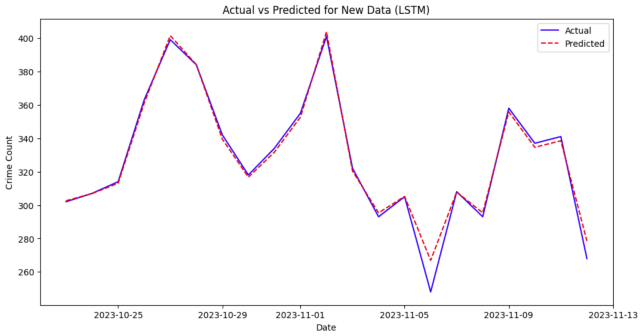
Finally we explored a bit of ARIMA and SARIMA. However, just like the XGBoost model, its performance

wasn't great with ARIMA Metrics: MAE: 41.81MSE: 2719.35, RMSE: 52.15. SARIMA Metrics: MAE: 32.22 MSE: 1470.42 RMSE: 38.35.

Model	MSE	MAE	RMSE
SARIMA	1470.42	32.22	38.35
ARIMA	2719.35	41.81	52.15

Now that we have all the model performances it is clear that LSTM performed the best. During the course of doing this and the main part of the project we realized since the Call for service dataset is constantly updating and we can use this to further evaluate the models. Up until now the prediction of the last 60 days and the datasets' last day was 22nd of October 2023(2023-10-22) and now as time passed we have new data ranging from 23rd October to 12th of November 2023. This was perfect as we could now further evaluate the models with new raw data. So we ran the models to predict the crime count for these days and plotted the actual crime count to compare the resulting graphs.

As expected LSTM performed the best,



the GRU performed decently well but not as good as LSTM and the performance of the other models were not noteworthy at all.

DISCUSSION

This project embarked on an ambitious journey to forecast crime occurrences in Tempe using a data-driven approach. Utilizing historical service call data spanning over three years, we aimed to develop predictive models that would not only offer insights into crime patterns but also assist in optimizing law enforcement resource allocation.

The models employed, including LSTM, Gradient Boosting Regressor, Feed Forward Neural Networks, and Random Forest, revealed intricate relationships between service call data and crime occurrences. Our findings underscored the significance of temporal factors, such as time of day and weekdays, in predicting crime rates. The analysis showed a marked increase in incidents during weekends, particularly on Fridays and Saturdays, highlighting the need for heightened vigilance during these times.

Comparatively, the Gradient Boosting Regressor emerged as the most effective model, especially when introducing the 'IsWeekday' bias variable. This enhancement led to a 3% improvement in the R2 score, signifying a better fit and predictive accuracy of the model. Such insights are invaluable for law enforcement agencies, enabling them to adopt more proactive strategies and potentially improving response times.

While our models achieved commendable predictive accuracy, the project faced certain limitations. The dataset primarily focused on service call data, omitting socio-economic, demographic, and other relevant factors that could impact crime rates. Future research could enrich the models by integrating these additional variables, offering a more nuanced understanding of crime dynamics.

Additionally, our daily-level analysis suggested that other variables like temperature might influence crime occurrences. Further studies could explore this avenue by incorporating environmental data to ascertain its impact on crime rates.

The practical implications of our findings are manifold. By accurately predicting crime hotspots, law enforcement agencies can optimize their resource allocation and response strategies. The ability to forecast crime occurrences based on data-driven insights marks a significant step towards more efficient and proactive urban governance.

In conclusion, this project represents a significant stride in the realm of crime forecasting. By harnessing advanced predictive modeling techniques and analyzing historical data, we have opened new avenues for enhancing public safety and law enforcement efficacy in urban areas. As the dataset continues to evolve with real-time updates, the potential for ongoing refinement and application of these models remains vast, promising even more robust and accurate forecasting in the future.

CONCLUSION

In our analysis, it has become apparent that the current dataset lacks information on various potential factors that could significantly impact crime rates. While we have made an earnest attempt to predict the number of incidents using the available data, it is clear that a more nuanced understanding of multiple influencing factors is essential to improve the accuracy of our model.

Some factors that are often associated with criminal behavior include socio-economic conditions, education levels, unemployment rates, and neighborhood characteristics. Additionally, demographic variables such as age, gender, and ethnicity can play a role. Factors like law enforcement presence, community programs, and the overall social environment also contribute to the complexity of criminal dynamics.

By delving into these factors more deeply, we can refine our model and enhance its predictive capabilities. This comprehensive approach will not only improve the model's accuracy but also provide a more insightful explanation of the variations observed in the target variable, ultimately contributing to a more effective understanding and prediction of criminal incidents.

FUTURE WORK

When examining the factors influencing the target variable in our daily-level data analysis, it becomes apparent that certain variables, such as the unemployment rate and inflation, may not exhibit significant changes on a daily basis. However, when transitioning to a monthly level, their impact could become more pronounced. In the current daily analysis, we are inclined to believe that temperature might be a pertinent factor. To delve deeper into this, we propose acquiring daily temperature data and establishing thresholds for categorizing days as either cold or hot. By integrating this information into our dataset, we aim to conduct additional experiments to validate and further elucidate the potential influence of temperature on the target variable.

REFERENCES

[1] C. -H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime ForG. Borowik, Z. M. Wawrzyniak and P. Cichosz, "Time series analysis for crime forecasting," 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 2018, pp. 1-10, doi: 10.1109/ICSENG.2018.8638179.

[2] G. Borowik, Z. M. Wawrzyniak, and P. Cichosz, "Time series analysis for crime forecasting," 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 2018, pp. 1-10, doi: 10.1109/ICSENG.2018.8638179.

[3] Sathyadevan, S., Devan, M. S., & Gangadharan, S. S. (2014). Crime analysis and prediction using data mining. In 2014 First international conference on networks & soft computing (ICNSC2014), pp. 406–412, doi: 10.1109/CNSC.2014.6906719

[4] Alghamdi, J., Al-Dala'in, T. (2023). Towards spatiotemporal crime events prediction. Multimed Tools Appl. doi: 10.1007/s11042-023-16188-x

[5] R. Yadav and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-5, doi: 10.1109/ICRAIE.2018.8710407.

[6] M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

[7] N. Jiang, K. Miao, Y. Chai, D. Lu and J. Wu, "Spatio-temporal prediction of crime based on Data Mining and LSTM network," 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2023, pp. 672-676, doi: 10.1109/ITNEC56291.2023.10081985.

[8] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.

[9] Biswas, Al Amin, and Sarnali Basak. "Forecasting the trends and patterns of crime in Bangladesh using the Machine Learning Model." 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2019, https://doi.org/10.1109/icct46177.2019.8969031.

[10] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.