Modelling the correlated factors with Victorian LGA offence rate, to guide government forecasting of offence count and rate in 2024

Group W12G7

Liyu Ren
COMP20008
liyur@student.unimelb.edu.au

Tianrui Qi
COMP20008
tianruiq@student.unimelb.edu.au

Luong An Khang
COMP20008
ankhangl@student.unimelb.edu.au

Executive Summary

This report integrated and normalized the dataset on median housing prices, electronic gaming machine (EGM) spendings, local government area (LGA) community features, and offence rates to create a standardized set of records about yearly crime rates and their relevant factors in different regions. Through correlation analysis, trend analysis and tree-based models, we have identified that the key factors which correlates to regional crime rates are poverty and socioeconomical indicators. Using machine learning techniques, we've identified both a support vector regressive (SVR) time series model for predicting regional crime rates and a linear regression model for regional population, both of which performed significantly better than the equivalent baseline model. Based on historical data, we've generated the upcoming 2024 offense count forecasts using both models and identified the most at-risk LGAs with the highest projected increase in offense counts.

Introduction

Crimes are unavoidable in the population of today. Usually, people view crime rates as a proxy for the quality of a city, rating high offence rates as undesirable features when deciding for vacations or residence; this report flips this concept around using data-oriented methods that find correlated factors to offence rates. Formally, we will explore a mixture of historical data on the Victorian local government areas (LGA), analyse it by deducing the factors that correlate with community crime rates, then model the relationship between these factors and the time series crime rates to forecast future crime rates and trends, finally identifying the at-risk cities. We hope our forecasts and deduced factors will help the Victorian State Government better comprehend the statistically likely reasons for high crime rates and direct their attentions to the cities potentially at-risk. More generally, it is possible to conduct further exploratory and designed studies on the conclusion correlations to gain a causational relationship that is valuable in improving the wellbeing of Victorians. The data sources we will use include: the "Victorian Communities" dataset containing demographic, geographic, economic, and accessibility information of suburbs in Victoria collected in 2011 and published in 2014; the "House by Suburb" dataset containing historical median housing prices of suburbs from the year 2013 to 2023; the "Electronic Gaming Machine" dataset with historical data on losses on electronic gaming machines across LGA in Victoria ranging from 2011 to 2023, collated between provided dataset and online sources (Victorian Gambling and Casino Control Commission); and lastly, "LGA Offences" dataset with crime rate statistics of Victorian LGAs collected by Victoria police ranging from 2014 to 2023.

Methodology

Our approach to prepare and preprocess the data is a mixture of data crawling on a list of LGAs and localities, data imputation on the ERP data, format normalization on the LGA column, data transformation and normalization on the communities dataset, LGA-grouping on the time series datasets, and finally joining between data for full feature sets.

In data analysis, we utilized Pearson's correlation coefficients and normalized mutual information to find correlated features with crime rate and population. To aggregate data by LGA for trend analysis, we've used weighted medians on the housing prices, appropriate weighted and normalized averages for community features, and simple averages for unweighted community features. For visualization, we've used pairwise scatterplots to identify correlated features towards populations and crime rates, and boxplots for both a graphical summary five number statistic on both the factors and the dependent variables, as well as for comparisons between factors over time or over regions.

For modelling crime rates, we've used various supervised models: decision trees for community-focused models and linear regression, support vector regression, and a multi-layered perceptron for the time-series models. Both encodes categorical features with one-hot encoding, and some non-linear models like SVR and perceptron have a normalization step. The community-based models are trained mainly on the features selected from community dataset (with feature selections discussed later), while the time series model adds the historical EGM and housing data. For time-series models, we've used the train-validation-test method: stratifying the 2023 records for the test set and using randomized 20-fold cross validation on the 2016-2022 train-validation set to train, tune hyperparameters and select the best models based on their RMSE. Communities-based models use the same method, but with 10% randomized test set, since the data is too limited for stratifying based on year. To derive future offence count, we've used a linear regression model for population modelling, and evaluated it with 20-folds cross validation. The model for both crime rates and population are retrained on the whole dataset for forecasting 2024 rates.

To generate the forecasts, we've multiplied the crime rates forecast and the population forecasts. To interpret the model, we've visualized the forecasted 2024 offense counts using a bar chart and performed analysis into model's trained parameters to reveal relationships between features and crime rate, as well as performing sensitivity analysis on the range of hyperparameters.

Preprocessing and Exploratory Data Analysis

Preprocessing and Data Cleaning

As we want to join the four separate datasets, it is necessary to clean and impute each set individually as to both generate the joining column and improve its data quality. All tables contain the information for LGA or communities within LGA, making it a sensible candidate for joining.

To start, we cleaned the LGA column present in three out of the four datasets (excluding housing). Firstly, we crawled an official list of the LGAs from the Victorian Government website (Victorian Election Commision, 2024) for validating the LGA names. We discovered that some datasets include the full LGAs naming, while others merely contained a shortened version. Our fix is to normalize all LGAs into a unique form with no whitespace, removed dashes, the latest 2024 naming (we've discovered that Moreland was changed to Merribek in 2023), and removal of all stop words like "of", "city", "rural", "shire", and "borough". We must drop a few rows with non-standard LGA labels, including "Unincorporated Vic" and "Justice Institutions and Immigration Facilities". Thankfully, these entries are rare with few offence counts.

For the EGM datasets are split between a 2011 to 2020 version and a 2021 to 2023 version (Victorian Gambling and Casino Control Commission, 2024), we need to preprocess the EGM datasets by combing the two in a spreadsheet program and aligning the years and LGAs. From the spreadsheets, we quickly noticed some annotations dictating that some LGAs are combined in the EGM to form "amalgamated" LGA, likely due to their small EGM expenditures. This prompted us to perform the same combinations on the LGAs rows in the offences, communities, and housing dataset to improve the data consistency. Sadly, this limited the range of LGAs for our data analysis and modelling to the relatively medium and large cities and towns. Additionally, we feature engineered an "EGM spending per person" column using the yearly population data from the offense table.

For Community, 103 top features that could affect crime rate are manually chosen from the full columns. For cleaning, all "<5" is converted to "5". Missing "2007 ERP age group" statistics are imputed using the mean values of communities within the same LGA. If that's not available, the values are roughly imputed with the 2012 data. After cleaning, we grouped all entries by LGA. Transformation/Derivation of features are required before grouping: number of Dwellings (derived through "occupied private dwellings" and corresponding %), Location's x and y coordinates (from "Location"), Population (from Density and Area), and converting % value to actual count for grouping (and rederived the % after).

There are 3 main types of aggregation methods on the features, depending on their semantics:

1. Summation: For attributes such as “Population”, “Number of Dwellings”, etc.
2. Unweighted average: For most attributes such as “Location”, “Travel time to GPO (minutes)”, etc.
3. Weighted average: For “ARIA+” and “IRSD” scores where the average is computed using population-weighted method (as specified in the “data definitions”).

After the grouping, to allow comparison between LGA and reducing noise from the obvious confounding factors such as Population, Area, Number of Families, etc., all relevant attributes are normalized with their respective main confounders. For example, the number of “Public Hospitals”, “Unemployed, persons” and “ERP age groups in persons” are normalized with “Population”, while number of “Occupied private dwellings” are normalized with “Number of Dwellings”. The “ABS remoteness category” is rederived from “ARIA+ (avg)” using the expert thresholds specified in data definitions, and ordinal encoded (0: Major Cities to 4: Very Remote).

For Housing, each entry is identified by Locality (Suburb), requiring conversion to LGA. The Locality list for each LGA was crawled from Vic Government Website (Victorian Government, 2024) to map all locality to its corresponding LGA. Due to typo and outdated information from the website, 34 Localities were manually mapped with Google Search. After converting all Locality to their corresponding LGA, we grouped all entries by their LGA and derived the population-weighted average for the housing price. The population for each Locality is collected from the Community dataset. If the population value is missing, it was imputed with the average population from other Localities with the same LGA. If all localities in an LGA are missing from Community dataset, we assume they have the same population and use the unweighted average housing price.

To join all the datasets together, we’ve converted each data frame into its separate table in a relational Sqlite3 database for ease of visualization. This structured format helped us to decide on a row format indexed by LGA and Year, containing two time series entries on the EGM spendings and median housing price, and a static set of descriptive community features indexed on LGAs only. Not only is this structure easy to generate from outer joining the tables, but it also incorporates most relevant features we need for later analysis.

Data Exploration and Visualization

The overall shapes for all Communities features are unimodal, bell shaped and often skewed with a few outliers due to the major urban regions. Since there are +100 columns in the Communities dataset, we only analysed the most important features with correlation to population and crime rate. For population analysis, Figure 1 (right) compares the population distribution for 2007 and 2012 from the “communities” dataset, which shows a similar distribution between the two years. The medians are almost identical, with a slight decrease in 2012. The distribution ranges and IQR ranges for both distributions are similar and there are no outliers. Overall, the total population has not changed significantly across the five-year period, with consistent variation between 2007 and 2012. Figure 1 (left) further explores the population trend for each LGA from 2014 to 2023 - approximately linear, with a dip from 2020 to 2022.

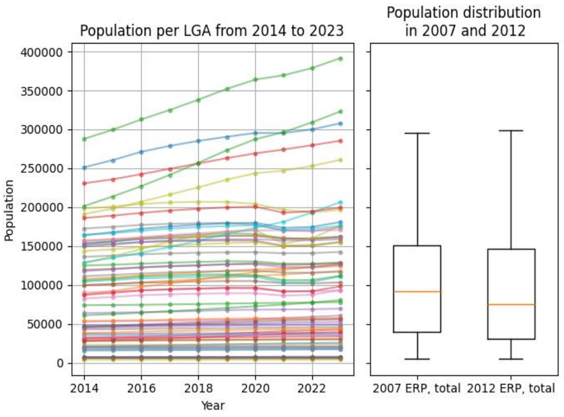


Figure 1: Population trend per LGA on the left. Box plot of population distribution for 2007 and 2012 on the right.

The histogram in figure 2 is heavily skewed to the left, with most of the LGA’s having a low population density of less than 500 people per kilometre squared, which includes more than 35 LGAs. Only a few LGAs have higher population density. The two outliers are Stonnington and Yarra, with population density of around 4000 persons per kilometre squared. The finding is consistent with expectations as

LGA's with high population are clustered around Melbourne, and Stonnington and Yarra are both in the heart of Melbourne. On the other hand, LGAs with smaller population densities are on average further away from Melbourne. The box plot in figure 2 further shows the skewness in the population distribution, with most LGA having low population densities, and a few outliers. This highlights differences between highly urbanised areas and less densely populated regions.

In searching for potential related features for population modelling, relationship between population and others are investigated. The Pearson correlation of population against EGM loss is 0.0235, and 0.3326 against housing prices. Both implies no linear relationship. However, Figure 3 shows a strong negative correlation between population density and travel time to nearest hospital, as population density increases, the travel time to the nearest public hospital decreases. Potential reasons for this correlation include that higher populated regions have greater economic activity, which implies greater development of infrastructure to meet the demand more effectively. The reverse may also be true. A more developed region can attract more residents due to easier access to public services, which increases the region's population density. The Spearman correlation is calculated to be -0.758814, which suggests a consistent inverse relationship between population density and travel time to the nearest hospital, despite the skewness in the data.

The three time-series dataset requires a different analysis compared to than the other features for their autoregressive time nature. For house prices, figure 4 displays the median housing price across all LGAs in Victoria from the year 2013 to year 2023.

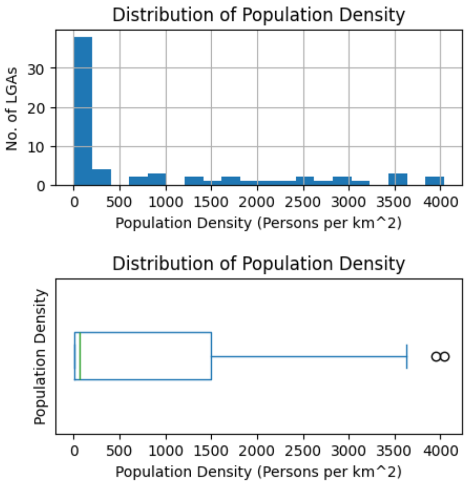


Figure 2: Histogram on the left showing the distribution of LGAs based on population density. Box plot on the right showing the central tendency and spread of population density across all LGAs.

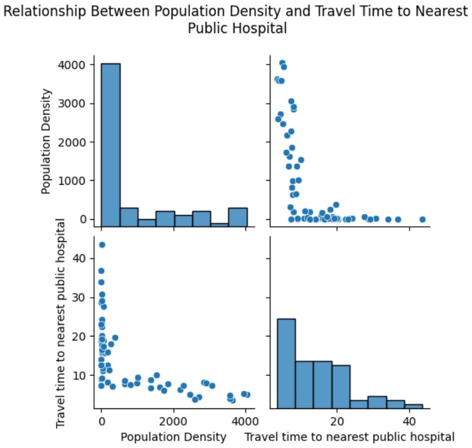


Figure 3: Pair plot showing the relationship between Population Density and Travel Time to the Nearest Public Hospital.

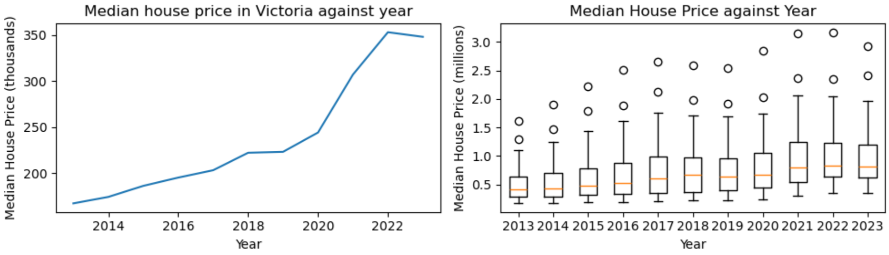


Figure 4: Line plot and boxplots of the grand and per LGA median house price in Victoria by year. Grand median calculated using weighted median method by population

From the left subplot, we can see a clear linear increasing trend of the median house price over the years, with a sudden increase in 2020-2021. The right boxplot showcased a positive skewed distribution of median house prices by LGA across all years, and two significant outliers every year: Stonnington and Boroondara both to the east of Melbourne Central. Upon reviewing the two regions, the outliers are explained by their proximity to the Melbourne CBD, which is a desirable property that increases the demand for housing and therefore its price. Overall, the grand median trend matches our expectations of a stable linear relationship for its likely dependence on the linear growing population size and capturing economic events such as COVID-19 in 2020. The positively skewed median LGA prices are also expected from its minimal dependency on skewed populations. Interestingly, the shape of each boxplot over the years remains fixed after scaling, this suggests that price changes in the housing market affects all LGAs relatively equally, with no significant preferences between the LGAs.

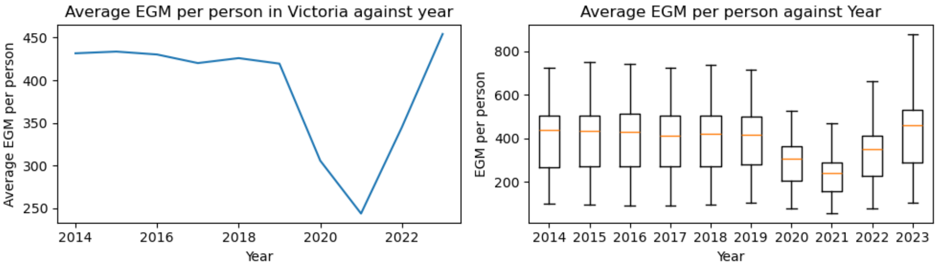


Figure 5: Line plot and boxplots of the grand and per LGA mean EGM spendings per person in Victoria by year. Grand mean is population weighted.

For EGM spendings, figure 5 displays their statistics ranging from 2014 to 2023. The left subplot displays a relatively stable average EGM spending per person over the years, barring 2020 and 2021. The dataset attributes the sudden drop in 2020 and 2021 to COVID-19, when casinos and gaming services were closed due to lockdowns. We decided to include this outlier period in the final dataset, for it reflects the economic and social conditions experienced during the pandemic years that can also influence crime rates. The right subplot shows a similar stable trend of EGM spendings by LGA over the decade. Importantly, the distribution of EGM spendings per person by LGAs is slightly right skewed, but rankings of EGM spendings per person between LGAs looks to be unchanged between years judging by the similarity in shape of the yearly boxplots, and it can be observed that macroeconomic impacts to EGM spendings mostly affect all area spendings relatively equally.

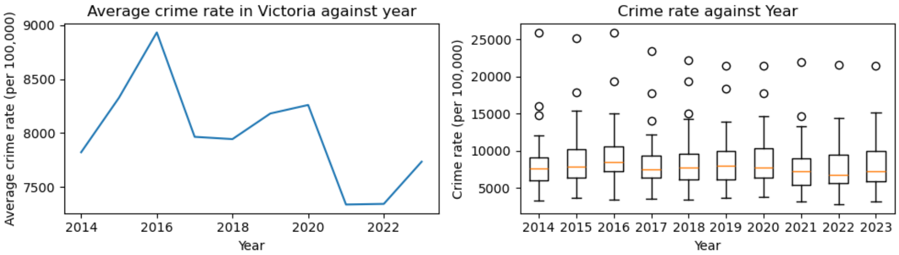


Figure 6: Line plot and boxplots of the grand and per LGA mean offence rate in Victoria by year. Grand mean is population weighted.

For crime rates, figure 6 displays their trends over 2014 to 2023. The left subplot shows an approximately mean reverting average offence rate that oscillates around 8,000 counts every 100,000 people per year. There is a peak average crime rate of above 8800 in 2016, and a low in 2021 to 2022 of around 7300. It is observed that increases in median house prices in 2021 correlates with the decrease in crime rates of the same period, and that a decrease in EGM spendings is related to a drop in crime rates also during 2021. However, these relationships are not perfectly clear and may not hold for all years. Nevertheless, this suggests that the crime rate model should include these. The right subplot shows a different trend, with each LGA experiencing a different crime rate trend over time. For instance, the highest outlier, Melbourne, did not experience a sudden drop in crime rates over the pandemic, while the second outlier, La Trobe, has a typical trend that reflects the grand mean rates. This suggests that there are LGA specific factors related to their crime rate patterns that we must include in a model. The boxplots also show a relatively symmetric distribution of crime rates between the LGAs when excluding the three outliers, seen by the equal length whiskers and boxes.

Crime-rate Correlation analysis

Melbourne is an outlier in offence rate, thus it'll generally be removed to reveal the general trend. For Communities, due to incompatible time, all correlations are analysed against the offence rate in 2014.

For linear correlation, only 4 features yield significant coefficients above 0.55, with valid trend verified through the scatterplots. There is vast difference between features, with the maximum coefficients at 0.80, and minimum at 0.30. With the high linear correlation, linear regression is a valid candidate model.

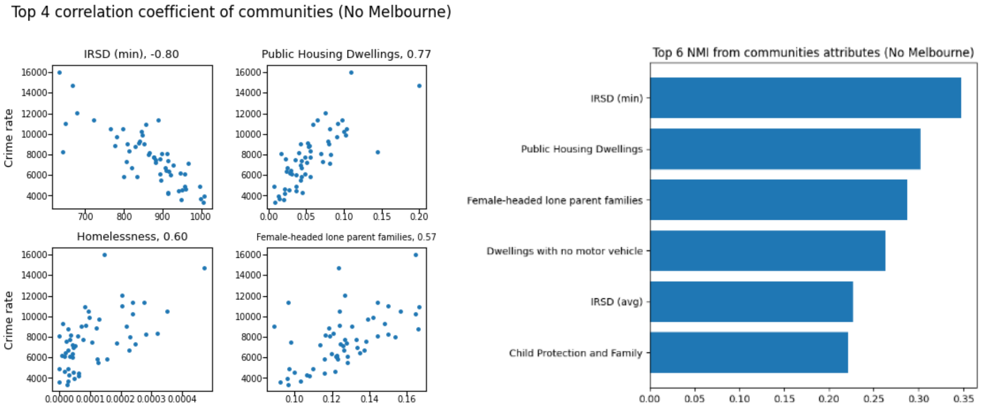


Figure 7: Scatterplot of top attributes of communities against crime rate and bar chart of NMI scores

To compute NMI, all columns (except discrete "ABS remoteness category") are discretised using equal-frequency KBins ordinal encoding (default 5 bins). Shown in figure 7, the maximum NMI is only 0.35 (and the minimum of 0.1), so decision tree classifier might not be a good model, although more advanced binning techniques can improve results, and different metrics than NMI can impact performance.

For EGM Loss (Figure 8), to predict the crime rate at time t, EGM loss of the same year is not available to use as a feature. Thus, the analysis is between the EGM loss at (t – shift) against crime rate at t.

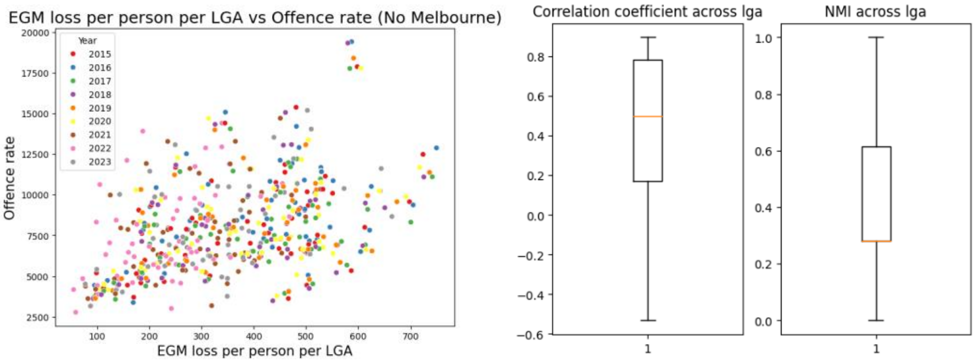


Figure 8: Scatterplot of EGM loss rate against crime rate, and Box plot of Correlation coefficient and NMI across LGA

The total EGM loss must be converted to loss rate per person to be comparable with the crime rate, avoiding population being a confounder. We normalized the EGM loss using the derived population from the Offence table (Offence count/Offence rate). At shift = 1, through the scatterplot and the correlation coefficient of 0.4251129, no strong overall linear trend exists between loss rate and offence rate. Overall, NMI is quite low at 0.08248. LGA might be a confounder, thus we analysed the scoring across LGA. The correlation coefficients across all LGA varies between the 2 extremes (-0.6 and 0.8), with the median near 0.5, indicating an overall weak positive linear trend. The NMI for each LGA varies widely between 0 and 1, with the median near 0.3, showing no significant information gains. Still, the score is much higher than grouping all points, suggesting the relationship is clearer for each LGA. Overall, the EGM Loss at shift 1 might be a decent feature for linear regression, while it might not be favourable for decision tree classifier using NMI. The higher the shifting in time, the weaker the correlation.

For Housing Price (Figure 9), due to same reason as EGM, the analysis will focus on the correlation between Housing price at time (t – shift) against crime rate at time t. For shift = 1, no clear linear trend between the Housing price and offence rate (coefficient = -0.08) or non-linear correlation (0.07).

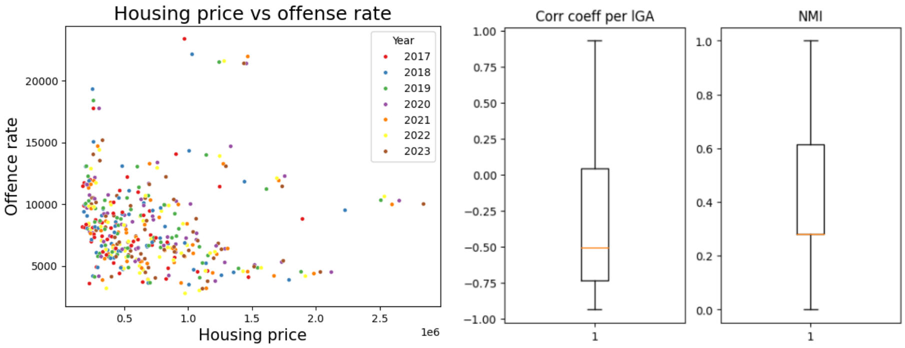


Figure 9: Scatterplot of Housing Price against crime rate, and Box plot of Correlation coefficient and NMI across LGA

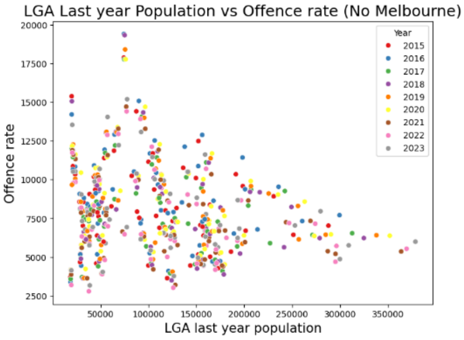


Figure 10: Scatterplot of last year population against offence rate

Checking the LGA, the scores vary significantly (from –1 to 1 for linear correlation, and 0 to 1 for NMI) with minimal median (-0.5 for correlation coefficient and 0.3 for NMI), indicating each LGA has a different trend with varying strength, yet weak overall. Thus, LGA is an important confounder, but otherwise Housing Price doesn't show significant correlation with crime rate at shift = 1. The higher the shift, the weaker the correlation.

For Population, since we can derive the population per LGA per year from the Offence table (Offence count/Offence rate), we quickly checked if there is any correlation with offence rate. From the scatterplot, low correlation coefficient at -0.21326, and low NMI at 0.04679, there is no significant correlation.

Data Modelling

We decided to model the population and crime rates separately. Population is separated into another model, since it is one main determinant of the offence count, and relatively easy to forecast. By removing such significant determinant, we hoped that crime rate becomes more stabilized and easier to model than offence count. Combining both, we can still forecast both crime rate and offence count.

The effectiveness of the crime-rate predictive models is determined by RMSE for *all models* (with comparison to a baseline model), due to Regression task nature, easy interpretation (since RMSE has same unit as the prediction feature) and relatively few outliers (based on EDA). Due to the differences between LGA, the common baseline model using the means of training crime-rate as prediction leads to extreme error (x5 compared to our proposed baseline), thus meaningless comparisons. Rather, a simple, yet more meaningful, baseline model is proposed: Using last year's crime rate as the prediction.

Crime-rate Models

These four models aim to predict the crime rate one year forward. All models one-hot encode the LGA to enable the models for learning personalized trend in each LGA (due to differing crime rate average and trend).

Community-focused models are developed using Communities as main features, for its rich, correlated features. However, due to extremely limited data, those models could not be used to forecast 2024 crime rate, only investigated as a potential framework to forecast with more up-to-date Communities data.

Time series-based models focus on time-series tables, thus able to have a larger dataset, from 2014 – 2023 for 59 LGAs, and capable of forecasting 2024 crime rate. For features, to predict crime rate at time t, all models use the offence rate at (t-1) and (t-2), housing price at (t-1), (t-2), (t-3), and EGM Loss rate at (t-1), (t-2). Communities features are also included, but generally do not have significant impacts on the models due to being far out-dated to correlate with crime rate in 2020 and later.

Decision Tree and Forest-based Model (Communities-focused)

Decision Trees are the simplest model capable of capturing non-linear trends, essential for the complex 103 features in Communities' dataset. The existence of ordinal categorical such as "ABS Remoteness

Category", and the intuitive implications of potential hidden thresholds factors (such as unemployment level through unemployed rate) further motivates the tree-based models. Since crime rate is continuous, the prediction is a Regression task, requiring a Regressor tree. Ensemble forest models are utilized to further performance and avoid over-fitting, by taking the average predictions of multiple low-depth trees iteratively built to correct its predecessors' errors (AdaBoost). Additionally, decision tree offers a human-understandable selection mechanism for insights into the features' relationship with crime rate.

Since the Communities' dataset only consists of data in roughly 2014, we assumed the normalized attributes of communities would stay roughly constant until 2016, since the number of hospitals, etc. likely not vary widely within 3 years, allowing pairing with 2014-2016 offence rates.

Four sets of features are investigated (all selected in training steps):

1. Feature 1: All features from Communities, housing price, egm, offence rate at (t-1)
2. Feature 2: Feature 1, but with top 20 features with highest NMI from communities
3. Feature 3 and 4: Feature 1 and 2 combined with the PCA on all Communities features

The explicit PCA is included, since the Decision Tree cannot learn linear combinations directly like Linear Regression, and complex (deep) tree will overfit due to small size of available datasets. All Tree models are built using sklearn, with the automatic hyper-parameter tuning using GridSearchCV on training data, with potential values listed here: DecisionTreeRegressor(max_depth=[3, 5, 7]) and AdaBoostRegressor(max_depth=[3, 5, 7], n_estimators=[200, 300, 400])

Table 1: All models RMSE on 10-fold cross-validation on Training data for model selection.

Models/Features	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**
Baseline	839.65	839.65	839.65	839.65
DecisionTreeRegressor	1047.05	822.77	987.25	947.11
AdaBoostRegressor	693.65	706.40	723.52	700.27

Only AdaBoostRegressor consistently out-performs the Baseline model, with the best performance using Feature 1. On test set, AdaBoost + Feature 1 yields RMSE 100 (compared to Baseline model: 840.6). Due to limited test set (+10 samples), the performance might be over-estimated, but along with validation result, it confirms the superior performance of AdaBoost to Baseline model.

Linear Regression Model (Time series-based)

Due to the gradual change in crime rate, the current crime rate is quite strongly correlated with last year crime rate (as also highlighted with the Communities-based models), showing a correlation coefficient of 0.9691 on training data. Thus, a Linear Regression forecasting crime rate at time t, using previous years' crime rate, along with Housing Price and EGM Loss, is the simplest, yet promising, model, using the default features above. Implemented with sklearn. No hyper-parameters included.

Multi-layered Perceptron Model (Time series-based)

The multi-layered perceptron model consists of a normalization scaler and a feedforward neural network that outputs crime rate predictions. It is motived by the large amounts of features (of over 100 columns) and the decently medium dataset size (of above 300 rows), in hopes that the non-linear model can capture even the smallest relationship between crime rate and the features.

A normalization step is also performed to clip the data between the range of zero to one, as to prevent the gradient from exploding during training. The perceptron is implemented in PyTorch consisting of two hidden layers each with 16 nodes, over 100 input nodes containing preprocessed features for every LGA and every year, and a single output node representing the predicted normalized crime rate. The activation function is ReLU. The network is trained by the Adam optimizer with low learning rate of 3e-5 and used the Huber loss function, which is more suitable for regression tasks for it smooths initial gradient descent and is less sensitive to training outliers than MSE. All the hyperparameters are chosen from experience to not overfit given the small dataset. Overall, training is done over 200 epochs, and the final model is used to cross validate against the test set. The limit 200 was chosen as we've observed overfitting beyond this value using validation set.

Support Vector Regression (SVR) Model (Time series-based)

The SVR model consists of a normalization scaler and a SVM regressor with a non-linear kernel (chosen rdf as the default non-linear) that outputs crime rate predictions. It is explored after observing the non-linear nature of the crime rate model, and the relatively poor performance of the perceptron model. The SVM is a generalized regularized regression model; we believe that it can achieve better performance than the neural network due to its relative simplicity with fewer hyperparameters for tuning.

Since the SVM is not scale invariant, preliminary tests shown that a normalization scaler on its inputs performs the best, hence it is also added to the pipeline. The crime rates are divided by a conservative upperbound of 20,000 to prevent the unseen testing data to have a normalized crime rate above 1. The implementation uses the SVR model from sklearn, in addition to a hyperparameter grid search module that selects the best parameters using 5-fold cross validation. The regularization parameter C is set very high to around 30K to allow for flexibility in the crime rate relationship, and to avoid overfit, the epsilon is set quite high around 200 to increase the model's error tolerance.

Crime-rate Models comparison

Table 2: RMSE of the crime rate models on 20-fold cross validation.

Model	Baseline	Linear Regression	Multi-layered Perceptron	SVR
RMSE	873.51	754.85	750.25	702.53

The SVR model has the lowest relative RMSE, and thus it would be used for testing and generating the 2024 crime rate predictions. Its RMSE on the testing 2023 set is 694.39 compared to the baseline model's 783.78, which is a good performance on unseen data and shows its explanatory power.

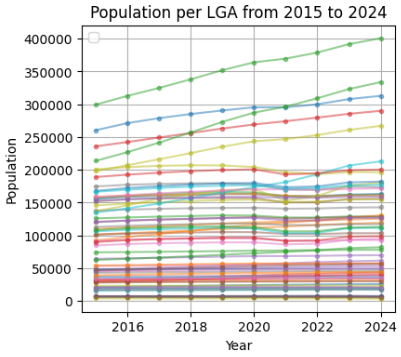
Population Linear Regression Model

To predict the regional population in 2024 to calculate offence count, we tried the simplest model using Year, LGA and the Last year's population to fit a regression line through time. This allows the model to enhance predictions using last year's population (since the variance with next year should be small) and utilising one-hot encoded LGA allowed adaptation for prediction on each individual LGA.

For evaluation, 20-folds cross validation is used, with comparison to Baseline model. The regression model has a low RMSE at 1900.680, at only 54% of the error of the Baseline model (RMSE: 3477.951). The results show the model performs well and indicates the feature set is likely to be optimal. This result is supported by our EDA, where features from other datasets are shown to be incompatible. Particularly, EGM loss and Housing price have low correlation (0.024 and 0.33), thus likely not be suitable features. Although Communities dataset has strong correlated features with population (like "2012 ERP total" at +0.8), the minimal data at around 2014 makes it too out-dated for the 2024 forecast. Therefore, using time-series data was the right choice.

2024 Forecast

Figure 11 and figure 12 shows the forecasted 2024 population and offense counts.



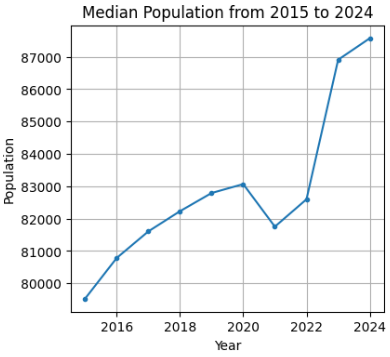


Figure 11: The time series graph on the left shows the historic and predicted population for each LGA from 2015 to 2024. The time series graph on the right shows the median population across all LGAs from 2015 to 2024.

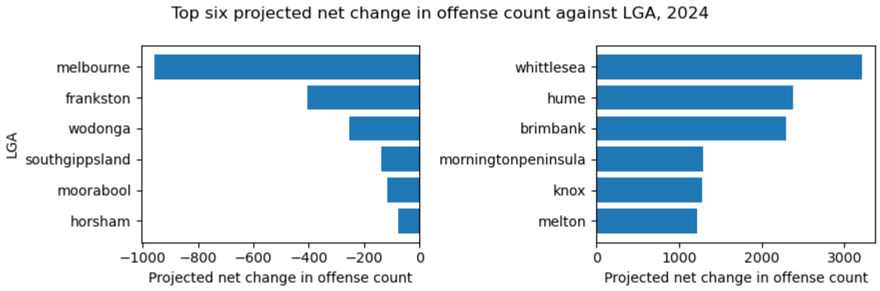


Figure 12: Bar plot of the projected net change in LGA offence count (using predicted rate and population) for the year 2024.

Discussion and Interpretation

For data analysis, the population trend for each LGA approximately follows a straight line, with a small dip from 2020 to 2022, and increasing again in 2023. This is most likely due to COVID-19. Most LGAs have similar and slow population growth rate, except for a few LGA in urban areas with higher population growth. With a growing population across LGAs, if the offence rate remains constant, the offence count will increase. Thus, this creates pressure for the government to reduce crime rate at a faster rate than population growth to keep offence count in check.

From trend analysis, both EGM spendings per person and median housing prices are only weakly related to crime rates when averaged across all LGAs on a yearly basis, with the relationships changing across years. It seems to contradict with popular sayings that higher average housing prices contribute to more crimes, or that higher average rates of gambling spendings have higher crime rates. However, this potentially could be due to the averaging across LGAs. The LGA Melbourne is also an extreme outlier in terms of crime rates and its surrounding LGAs are outliers for median housing prices and EGM spendings. This is likely due to a weak confounding factor on high population.

From correlation Analysis, for Communities, although the correlation coefficients are more significant than NMI, the top features are the same, including IRSD (min), Dwellings with no motor vehicle, Homelessness, etc., which are all indicators of socioeconomic disadvantage and poverty, thus highlighting the potential reduction of offence rate through socioeconomic support and improvement. For EGM Loss and HousingPrice, EGM Loss is more correlated with offence rate than HousingPrice. However, since both correlations are weaker than expected (neither > 0.5), common perception of low crime rate in wealthy neighbourhood (high housing Price) and low EGM loss region might require closer investigation. The sign of the coefficient, still, agrees with the common perception. Additionally, the correlation between EGM loss and Housing Price varies widely, thus suggesting each LGA possesses different trend/characteristics, emphasizing the needs to consider LGA as a feature in predictive models, and personalized government treatment for each LGA in combating crime rate.

The population linear regression yields strong performance. However, there is an overall sudden decrease in population across LGAs between 2020 and 2022, with population growth recovering in 2023. This is likely due to the COVID-19 pandemic, which demonstrates the limitations of a simple linear model when there are external factors affecting the linearity of the data. The government could provide additional data on external factors such as economic conditions, resident migration patterns, and infrastructure development data, essentially Communities data, but across multiple consecutive years to further capture trends in the population, which will improve the accuracy of offence count prediction.

For crime rates, the perceptron model performed slightly better than the linear regression model. However, it had several drawbacks in a lack of explainability for the effect of each feature and longer training time. Moreover, it may be difficult for the government to justify using the perceptron model forecasts for decision making as it is hard to support its outputs with logical reasoning to deduce suitable government policies.

The support vector regression (SVR) model performed the best with the fewest hyperparameters. Importantly, when tested, the model is resistant to slight changes in the hyperparameters. This likely signifies that it is a good model of the data that captured the important features of the training set. An insensitive model is important for the government when making its large policy decisions, because it can trust that their model is not sensitive to slight changes in assumptions or statistical omissions, saving funds in not needing to significantly readjust their existing plans from the previous forecasts.

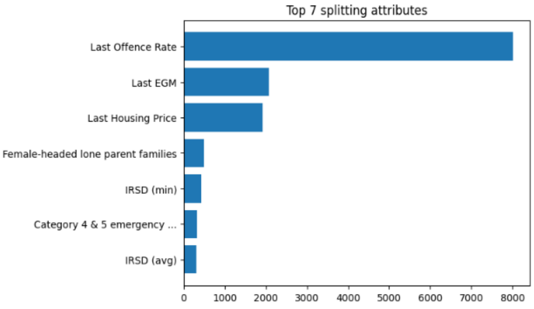


Figure 13: Bar chart of weighted total number of times used as splitting attributes

With tree-based models, we could identify the attributes' importance easily by counting the number of times each attribute is used as splitting attributes (weighted by the tree weights in AdaBoost "forest")

Shown in Figure 13, last year's offence rate is the key predictors, suggesting the gradual change in persisted crime rate, highlighting the need for long term management to significantly reduce the relative crime rate of an LGA. Since AdaBoost Model outperforms the Baseline models, other attributes have crucial connection and potentially worthwhile for government to keep track of sudden change. Surprisingly, Housing Price and EGM loss are ranked highly, despite minimal correlation during separate EDA, suggesting they might reveal hidden trends combined with other features. "Female-headed lone parent families" and "IRSD" are features for socio-economic disadvantage, and highly ranked, agreeing with the Correlation analysis. Although no confirmed causality, improvements to socio-economic disadvantages, such as support female single parents, likely reduce crime rates, through reduced incentives to commit crimes.

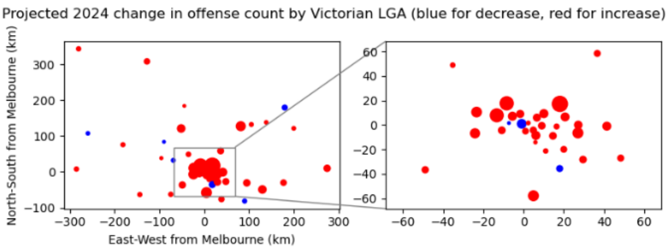


Figure 14: Geographic bubble plot of the projected net change in LGA crime counts for the year 2024

The 2024 prediction on change in crime rate (in Modelling) clearly highlights 49 LGA (out of 56) with forecasted increase in crime rate, requiring immediate attention. However, to justify the distribution of resources to tackle offences, the absolute count would be more beneficial, since the crime severity might be over-estimated in LGA with low population using crime rate. Thus, using our population model, we can predict the 2024 population to compute the predicted offence count in 2024. As shown in Figure 12, the changes in offence count vary, with Whittlesea require most attention (+2,000 offences) and Melbourne with the best prospect (-1,000 offences). With Figure 14, several regions with expected increase in offence count are clustered near Melbourne (suburb region), requiring the most resources/polices to tackle offences.

Limitations and Improvements

Communities.csv (roughly 2014): For each row about an LGA, its columns contain values from different years (2007-2014). We currently assume all attributes stay constant during the period for rows consistency and interpretability, thus each row is representative of 2014 statistics. There exist missing communities (such as Aintree). Moreover, as specified in "data definitions", random error is introduced to 2012 population, while 2007 population are only rough estimates, and some rows are removed due to inaccuracy (which we imputed with 2012 data), causing errors in prediction.

EGM.csv (2011-2013): Special "amalgamated" LGA defined, which require the similar grouping on all datasets, losing specific rows for individual LGA in the group. Only 69 LGAs are included out of 79. This limits the scope of our models and correlated factors.

HousingPrice.csv (2013-2023): Due to errors on VicGov websites, manual search for 34 localities is required. There's also confusing locality without definite LGAs like Coonans Hill, thus cannot guarantee accuracy.

Offences.csv (2014-2023): The offence rate for Melbourne is unusually high, so unclear if it's referring to either the City of Melbourne, the entire region of Greater Melbourne (multiple LGAs) or typo occurs.

All: Only 56 LGAs remain after cleaning and joining between all 4 tables, with around 20 LGAs missing. Even with supplemented EGM, the dataset only contains 400+ entries, quite small for complex models, like SVR, MLP and deep tree/forest models. This is likely to cause overfitting if the model has a lot of parameters.

Population model: The drop from 2020 to 2022 may negatively affect the performance of the model as it is assumed that the population trend is linear for all LGAs. Due to timing incompatibility, we could not build a population model using Communities features, but the correlations are still mentioned in EDA for future development on a better Communities dataset.

Tree-based model: Communities.csv only have data in roughly 2014, thus requiring the assumption of all attributes staying constant until 2016 to form a trainable dataset. Still, the dataset only has 111 entries over the 2 years, requiring us to select 10% as a randomized test set, instead of using 2023 as test set like other models.

Time-series model: Both the SVR and perceptron model required long training and cross validation times, making hyperparameter tuning difficult and tedious. The inclusion of past 2014 community features slightly improved their performance but is logically problematic, due to out-dated data.

For improvements, an updated Communities dataset with compatible time series from 2011 to 2023 will allow the usage of the rich features in all models, including the time-series model, without sketchy constant assumptions. This is likely to improve the prediction vastly according to the strong correlation shown in EDA. Expert clarification on missing LGAs and mapping of HousingPrice suburbs to their LGA will further improve the data quality.

Moreover, inclusion of EGM dataset without amalgamated LGAs will increase the scope of the models and findings to lower populated region, thus reducing the bias of forecast statistics and government interventions toward only larger LGAs.

Conclusion

Overall, the report fully answers the two key components of our research questions.

Identify correlated factors with the crime rates: Through EDA, supplemented with Tree-based models, the key features representing poverty and socioeconomic disadvantage, including “IRSD”, “Homelessness”, “Female-headed lone parent families”, shows the highest correlation with crime rate. Intuitively, these relationships are reasonable as poor socioeconomic conditions are likely to foster crimes. If a follow-up causational study confirms the relationship, it suggests a potential indirect approach to reduce crime rates via supporting socio-economic disadvantage people, through subsidies or anti-discrimination education. Referral to the attributes list can narrow the scope of most impactful actions, such as investing in low-rent housing for the homeless people, due to its high correlation (and roughly inferred causality).

Support forecast future crime rate and count: Several alternative models and feature-engineering techniques are employed, thus providing a comprehensive comparison and selection of the top model for Crime Rate predictions (SVR) and Population (Linear Regression), thus deriving the future Offence Counts. Both models outperform the Baseline models significantly, showing efficient use of data and concrete predictions in crime rate changes for 2024, with a full list of 49 “at-risk” LGAs expecting rise in crime rate at varying change, guiding an efficient allocation of resources prioritizing those regions. Further analysis also reveals the clustering of “at-risk" regions near Melbourne, suggesting the allocation of fundings/police forces in the centre of Melbourne for rapid mobilization to neighbourhoods.

References

Victorian Election Commision. (2024, October 3). *Local Councils*. Retrieved from https://www.vec.vic.gov.au/electoral-boundaries/local-councils

Victorian Gambling and Casino Control Commission. (n.d.). *Expenditure Data*. Retrieved from https://www.vgccc.vic.gov.au/resources/information-and-data/expenditure-data

Victorian Government. (2024, October 3). *Know you concil*. Retrieved from https://www.vic.gov.au/know-your-council