Capstone 2: US Mass Shooting Crime Modeling

Where & Why do the mass shooting accidents happen in the US?

Abstract

Building regression functions using linear (simple Linear, Lasso, Ridge, OLS) and ensemble (Random Forest) approaches by modeling variety data of US gun crime, mass shooting cases, gun licensing & state employment statistics. Targets of the study are pointing out where the mass shooting cases usually happen and revealing the potential reasons leading to such critical actions.

Random forest regression, which is achieved a R2 score of 77.17% with RMSE = 10.43 on the training set; And MAE = 10, RMSE = 21.38 on testing set, is selected to the best among the 5 approaches. Consequent of that, the key insights are emitted in order to rank to top 5 states where accidents happened the most and how other high level factors might effect to that behavior such as: shooter age, unemployment rate, accident seasonality, gun possession rate…And of course the high-level advice for FBI and US government on managing such risks to the community.

Key words: Mass Shooting, Gun Licenses, Mental Health, Machine Learning Model, Linear Regression, Random Forest, Feature important.

Introduction

Gun possession & abuse is a big controversial problem within US society nowadays. Therein, Mass shooting is a critical form of the gun crime.

It is necessary to reveal the mechanism causing such severe cases in the relationship with as many social features as possible, such as: gun possession rate, unemployment, mental health, gender, age…That helps the managements a foundation basis to issue & execute their gun moderator, management tools and optimize resource & allocation accordingly.

Methods

The study using 04 dataset: Total gun licenses per years; Total gun deaths per years; State employment statistic by years; And US mass shooting cases by years. These dataset have been loaded, cleaned, converted and cooked to be the standard formats then finally merged into a united dataframe of which the shape is 244x15 from a time period of 2009-03-10 to 2018-11-19.

15 features comprise: *'total_victims', 'age', 'employeed(Y/N)', 'employed_at', 'mental_health_issues', 'gender', 'year', 'month', 'monthday', 'weekday', 'state', 'licensed_business_entities', 'total_gun_deaths', 'state_population', 'unemployment_rate'. Therein 'total_victims' is target feature and the rest are independent variables.*

1. Exploratory data analysis

02 additional features have been aggregated for a deeper investigation:

'*sector*': is a categorical feature that classify shooter's occupation domain.

'*frequency*': is a continuous feature aggregated by counting total number of accidents per state from 2009 – 2018.

The Spearman coefficient (Figure1) between '*total_victims*' target to other features showing a low to moderate monotonic linear relationship, meanwhile some other pairs have a high to absolute high correlation values. It is initially supposed that linear model will be less efficient on explaining '*total_victims*' from others.

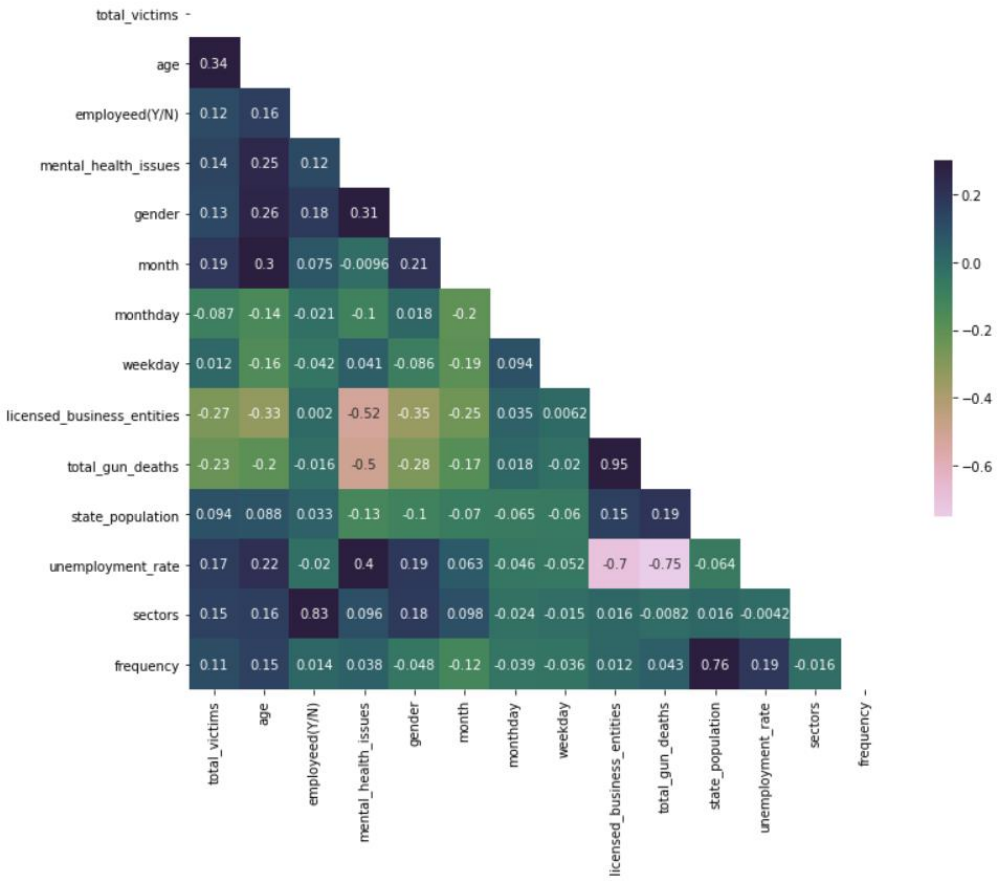

Figure 1: Spearman coefficient matrix showing a low to moderate monotonic linear relationship between 'total_victims' and the rest features.

The features then visualized by time from 2009 – 2918 showing a significant increase in number of total victims end of 2017. Suspect's dominant age range is from 15-45. And majority of shooters are male & jobless (Figure 2).

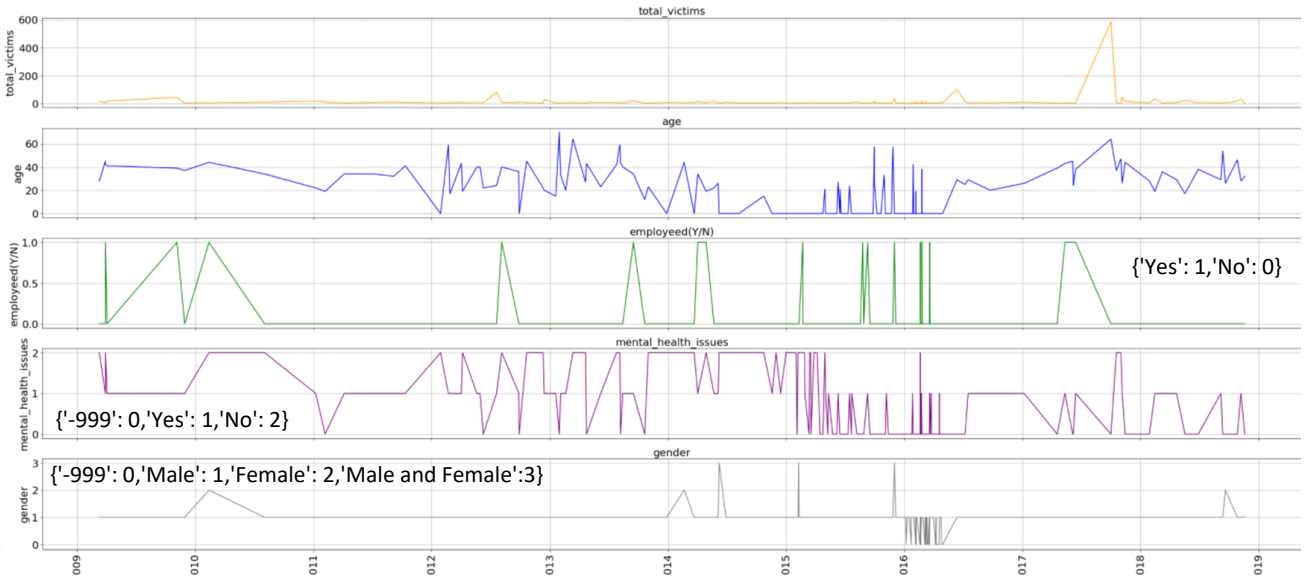

Figure 2: Time series visualization of shooter's age, gender, mental_health_issues, career status and total_victims features from 2009 – 2018.

There are strong positive relationships between *licensed_business_entities & total_gun_deaths*, yet fair to the *state_population* & weak to *unemployment_rate*. Companies/organizations where shooters used to work or being work spread on many sectors, yet the military contribute a great number (Figure 3).

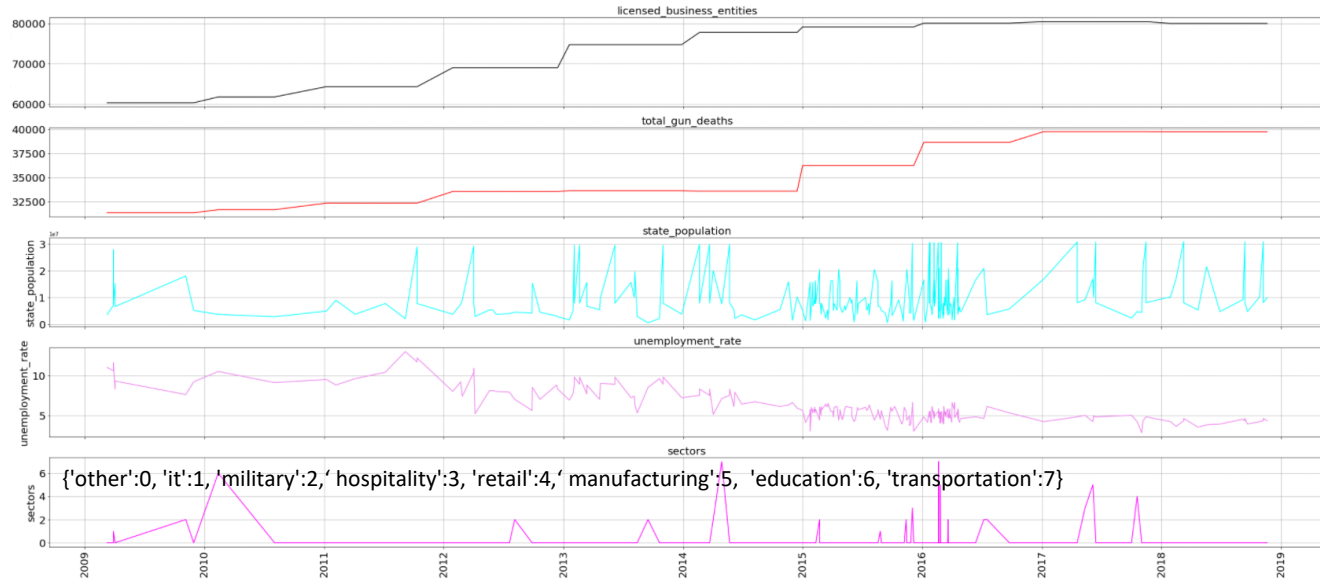

Figure 3: Time series visualization of licensed_business_entities, total_gun_deaths, state_population, unemployment_rate & sector feature from 2009 – 2018.

Such simple visualization above only extracts general info. In order to dig deeper for more value insights, the data were groupby by 'state' and plot in 3D dimension: *'state'*, *'frequency' & 'total_victims'* (Figure 4)

Top five states that have the highest number of mass shooting cases are: California, Florida & Texas New York

In term of severity level: Nevada is the state which has highest number of total victims, following is Florida & Colorado

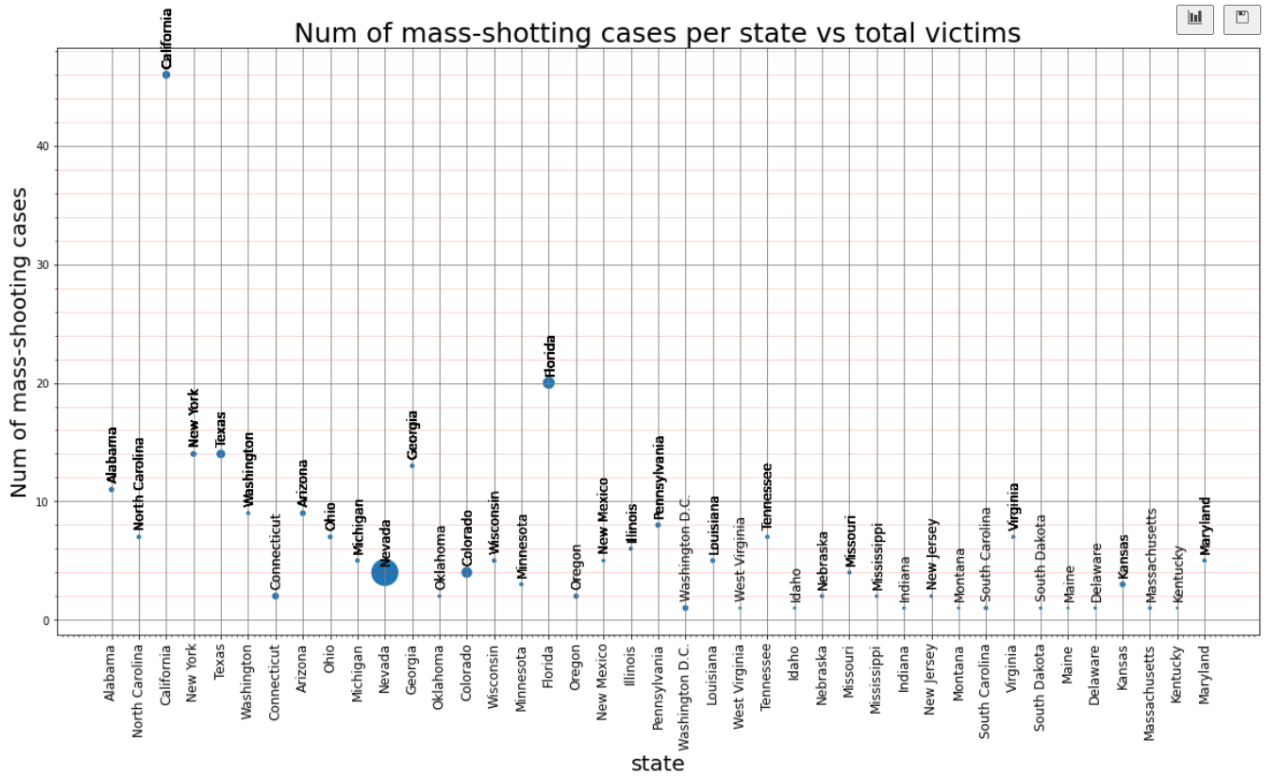

Figure 4: 3D dimension visualization: 'frequency' is Y axis, 'state' is X axis, circle size is 'total_victims'.

2. Data modeling:

'state' feature data type was still in object format, which is not suitable type feeding to model. It was converted to 41 one-hot encoding features (41 states). The final data shape was 244x58, it then chopped in to Y-target (total_victims) and X-independent features. Both X, Y were splitted to training & testing set follow the fraction 7/3, and ready for modeling step.

Linear & ensemble learning were both carried out to assess the key features contributing on explaining the target.

Key metrics that measure the quality of a regression model are:

- R-square on train set, the higher the better

- Mean Absolute Error (MAE), Root Mean Square Error (RMSE) on both train & test sets, the lower the better.
- Mean/Standard Deviation or MAE, RMSE for cross-validation (if applicable), the lower the better.
- Linear independent coefficient p-value less than 0.05 for a statistic significant.

Pipeline and GridSearchCV were applied to automate the modeling process and finding the best hyperparameters.

2.1. Linear regression Model:

As aforesaid at the very beginning of EDA, '*total_victims'* target has low to moderate linear relationship to other features, it is supposed for not a high quality model.

2.1.a. Simple linear regression

The best model hyperparameters was the number of best features (best_k) feeding to the model (Figure 5).

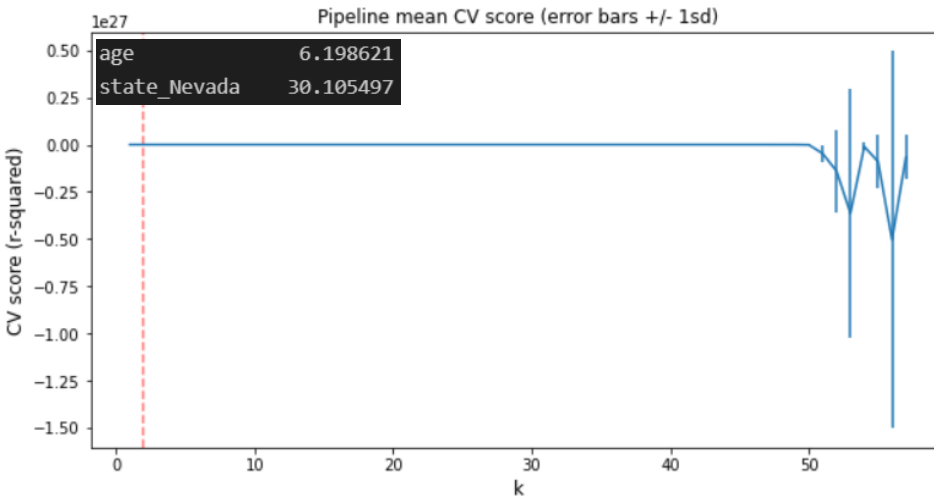

Figure 5: R-squared score on test set of simple linear regression model.

It is advised that the best_k = 2, they were *'age' & 'state_Nevada'* with the following metrics after cross validation 5 folds:

Mean_MAE = 11.9741, Mean_RMSE = 44.8658

Std_MAE = 6.9656, Std_RMSE = 44.2380

The model then ran with the recommended best_k, its outcome's metrics as following:

R-square = 0.4978, MAE = 14.7536, RMSE = 48.3944

2.1.b. Simple LR vs Lasso, Ridge regression

Advanced version of simple linear regression: Lasso & Ridge had also built with a range of alpha values. The model metrics as following:

lasso_params = {'alpha':[0.02, 0.024, 0.025, 0.026, 0.03]}
ridge_params = {'alpha':[200, 230, 250,265, 270, 275, 290, 300, 500]}

	LR	**Lasso**	**Ridge**
Mean R2_train	**0.5998**	**0.5969**	**0.1407**
Mean R2_test	**-8.9614**	**-8.6199**	**-0.4153**
Mean Rmse_train	**28.3077**	**28.4099**	**41.4781**
Mean Rmse_test	**47.4136**	**46.5938**	**17.8718**

It seems that Linear Regression & Lasso seem overfit, meanwhile Ridge seem underfit

2.1.c. OLS regression

The least square algorithm (OLS) was also used with the following metrics:

R2_train = 0.4926, R2_test = -2.0108, MAE_test = 14.7862

RMSE_train = 31.8737, RMSE_test = 26.0667

There are 04 features had p-value < 0.05:

Features	**P-value**
age	0.001178
weekday	0.013663
unemployment_rate	0.013958
age_group	0.011975

2.2. Ensemble Regression Model (Random forest regressor)

RF model had been run with data cross validation (5 folds), the best number of trees before taking the maximum voting or averages of predictions (n_estimators) was 143, following with the quantitative importance value of the features (Figure 6). The feature importance is an extremely value info for further insights investigation.

Cross validation metrics are:

Mean_MAE = 10.8882, Mean_RMSE = 10.8882, Std_MAE = 6.1939, Std_MRSE = 34.4004

Recommended most important features: *state_Nevada, date, age, state_population, monthday, month, unemployment_rate ,total_gun_death, license_business_entities.*

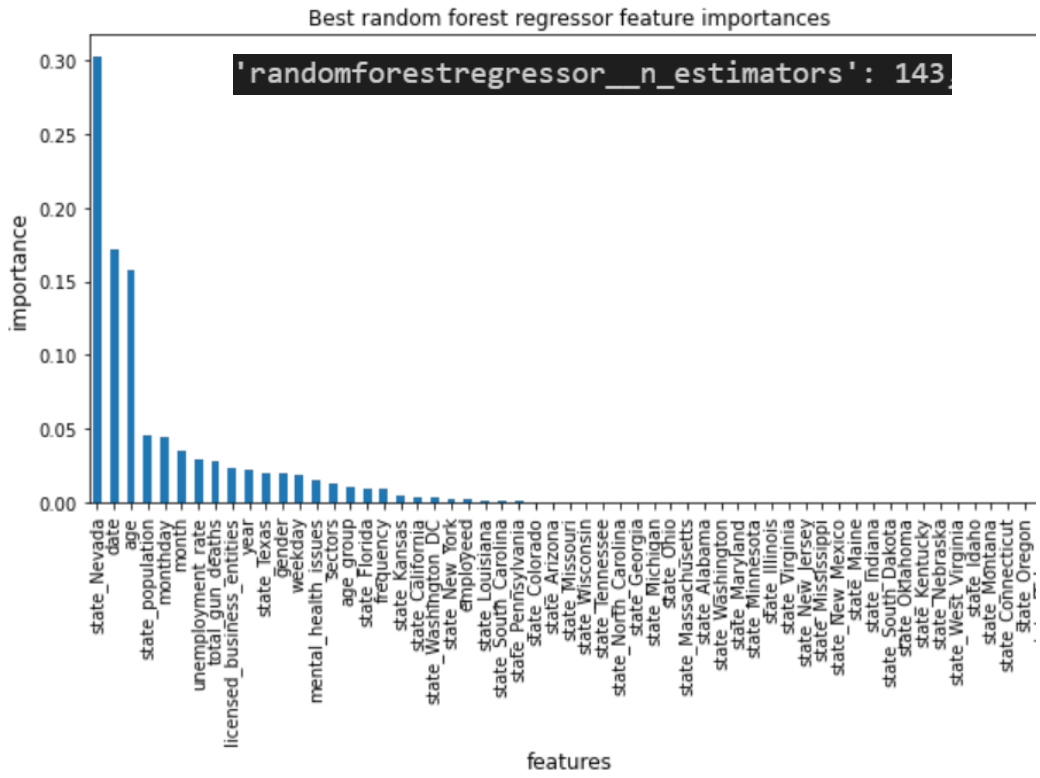

Figure 6: Feature importance of random forest regressor.

And the model metrics:

R2_train = 0.7601, R2_test = -3.1089 , MAE_test = 10.0009

RMSE_train = 21.9143, RMSE_test = 30.4516

3. Model evaluation

3 types of linear regression models have been conducted that all train_set R2 are less than 60%.

- Linear Regression seems working the best among them with train_set R2 = 59.9%, the RMSE in both train & test sets are not much different from Ridge. Numbers of best features are *state_Nevada & age*.

- The OLS showing train_set R2 = 37.5% with the statistic significant features are: *age, weekday, unemployment_rate.*

Random forest shows more promising result because of lowest MAE & RMSE values and the highest explained variance of 76% on train_set. 5 most important features are captures: *state_Nevada, date, age, state_population, monthday, month, unemployment_rate, total_gun_death, license_business_entities*

The final model was re-run with recommended features & optimum tree number (Figure 7)

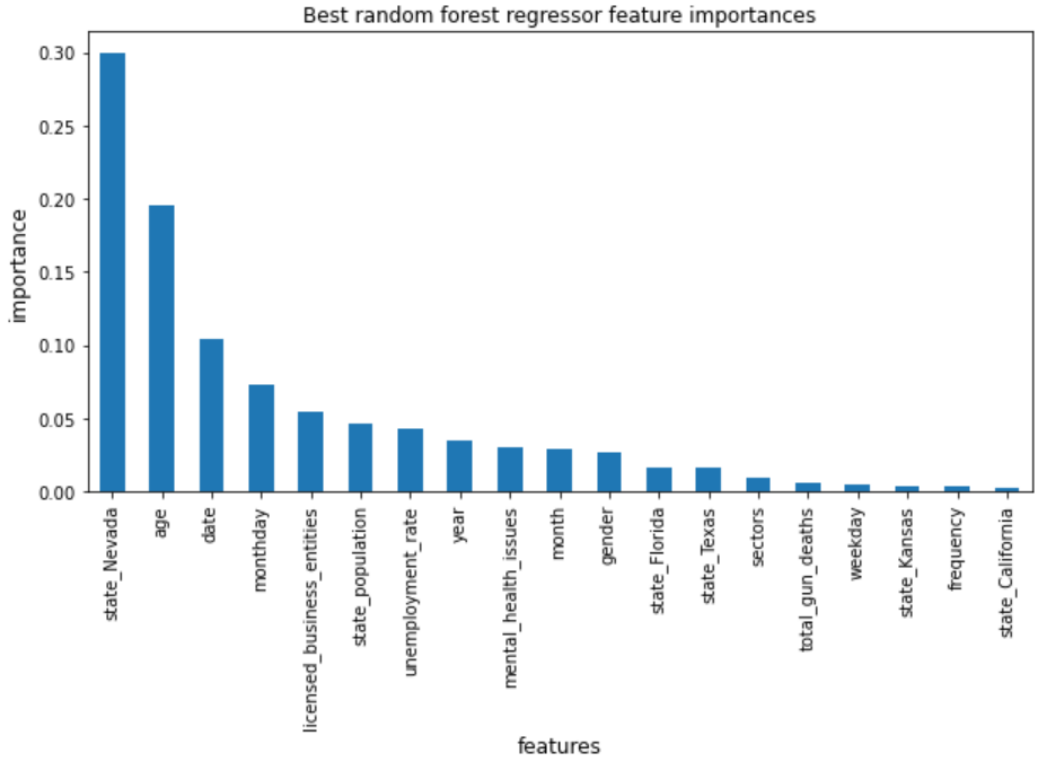

Figure 7: Final random forest regressor model

The key metrics as following:

R2_train = 0.7717, R2_test = -3.2392, MAE_test = 10.0009,

RMSE_train = 10.4332, RMSE_test = 21.3797

Results

Following the guidance from RF model outcome, more specific visualizations need to be added

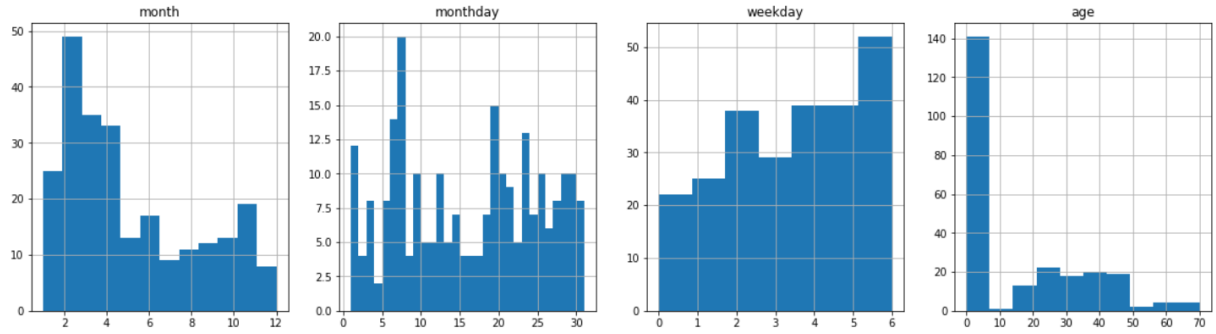

Figure 8: Histogram of month, monthday, weekday & age features

Mass shooting cases shows the high rate at the weekends & in spring. Moreover, shooter's age range drops in range of 15-49. Age range from 0-5 could be an outlier (Figure 8).

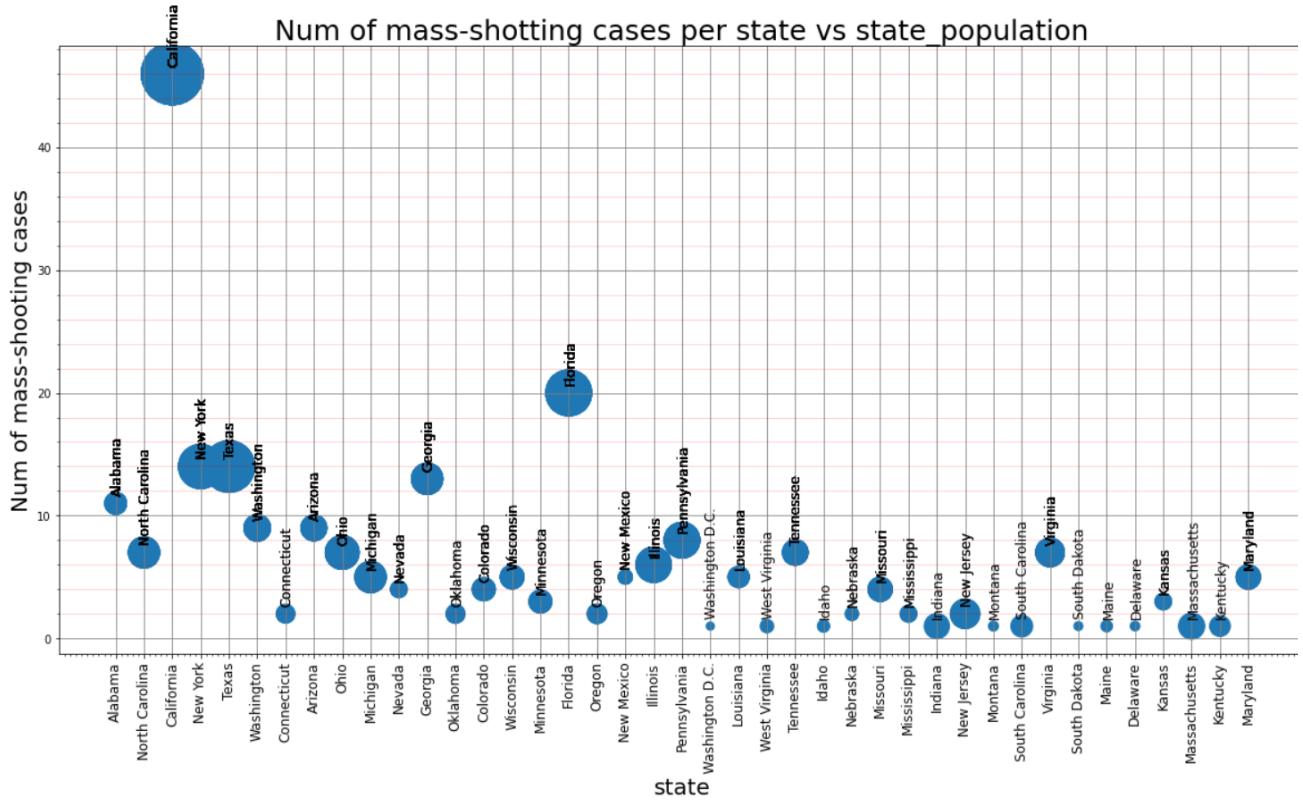

Figure 9: *3D dimension visualization: 'frequency' is Y axis, 'state' is X axis, circle size is 'state_population'*

The population scale seems having a positive relationship to number of mass-shooting cases. Top five states of which the highest population have also the highest accident occuration (Figure 9).

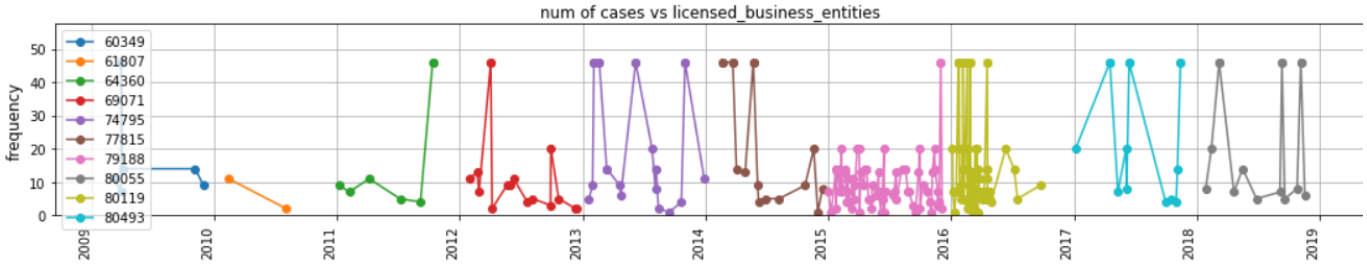

Figure 10: Number of cases vs gun licenses by year plot.

There is a positive relationship in number of gun licenses with number of mass shooting cases (Figure 10).

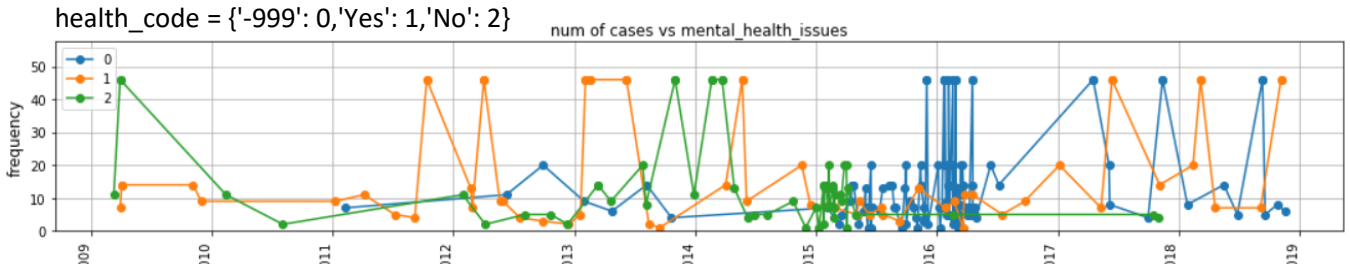

Figure 11: Number of cases vs shooter's mental health issues by year plot.

The number of mass shooting cases related to mental health issues seems increasing in the more recent years (2017 – 2018) (Figure 11).

Conclusion & Discussion

There is existing relationships between the *total_victims* and the set of features: *location, state population; shooter's age & mental health; unemployment_rate;* And *seasonality*.

There is evidence encouraging the government on improving employment rate & strictly managing gun licenses, and focusing to high population states.

Spend more effort on educating young people, promote more social connections & healthy activities, and maintain work-life balance.

The mass shooting cases show the high rate at the weekend & in spring.

The negative R2 on test set showing a fair model's quality. It needs to improve by feeding more powerful features, well-managing outliers and of course more data.

Random Forest model is selected to the best one for further investigation.