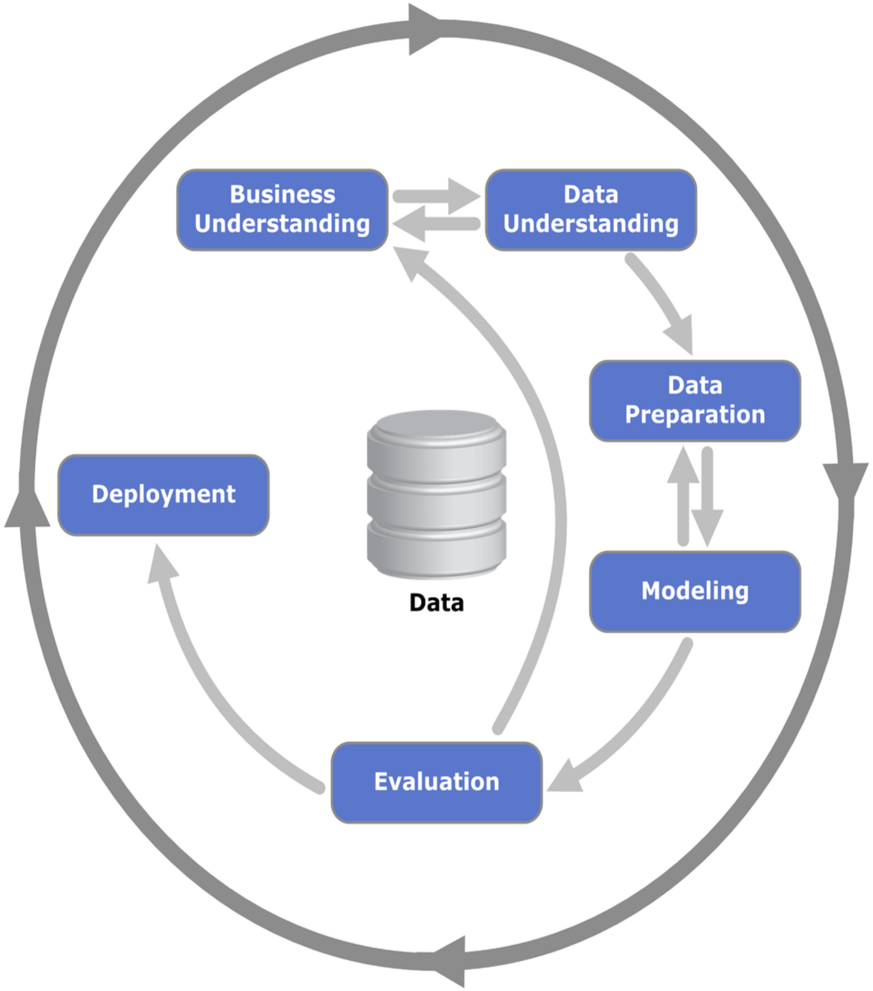
US Mass Shooting Crime Prediction

Springboard

Binh Kieu Nguyen
Madrid, Nov 07, 2021

Agenda



1. **Problem statements**
2. **Data wrangling & EDA**
3. **Features selection & modeling**
4. **Results & Conclusions**

By Kenneth Jensen - Own work based on:ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610

Problem Statement

Reveal related features from the mass shooting cases through US states, which could be the guidance for FBI and Government in policy revision. By exploring & analyzing FBI US mass shooting statistic & US gun possession dataset accompany with unemployment rate to build a predictive model of mass shooting cases in a context of population, unemployment rate, mental health, yearly gun possession.

1 Context

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

2 Criteria for success

Find out the relationship between mass shooting cases and key features: gun possession, unemployment rate, mental health, age, gender and others…to reveal the most important features contribute to the fatal cases then build a predictive model to predict the future cases.

Explainable features extraction
ML predictive model for future cases prediction.

3 Scope of solution space

US crime 1997 – 2016
Type of weapon involve by offender 2010 – 2019
US mass shooting 1966 – 2019
Staadata (population & unemployment rate)

4 Constraints within solution space

Solution in time bounded

5 Stakeholders to provide key insight

6 Key data sources

FBI crime dataset
Kaggle US gun database, US unemployment

Project US Mass Shooting Crime Prediction

Data Wrangling

Dataset:

1.GunPossession_1986_2018: Total of gun licenses per year, from 1986-2018
2.GunDeaths_2009_2018: Total gun deaths per year, from 2009-2018
3.Employment: State population, labor force & unemployment rate, from 1976-2019
4.USMassShooting19662019: Number of mass shooting cases accompany with total victims, time, location (state) & suspects' info (gender, age, employment status, employer, mental health issue), from 1966-2019

Data cleansing:

Remove duplicate & unnecessary columns
Extract states from locations, correct states 'name convention
Convert data type from object to numeric one

Data merging:

04 different data have been merged into a dataframe which has the shape of 244x22, from 2009-03-10 to 2018-11-19

Features comprise:

'fatalities', 'injured', 'total_victims', 'policeman_killed', 'age', 'employeed(Y/N)', 'employed_at', 'mental_health_issues', 'gender', 'year', 'month', 'monthday', 'weekday', 'state', 'total_licensees ', 'licensed_business_entities', 'population', 'total_gun_deaths', 'total_children_teen_gun_deaths', 'state_population', 'state_labor_force', 'unemployment_rate'

Project US Mass Shooting Crime Prediction

Exploration Data Analysis

Features definition:

Target: total_victims

Explainers: the rest features except total_victims

Features transforming:

Create a log scale of total_victims for visualization

Type transformation of suspects gender, mental_health_issue features to numeric ones

Feature extraction:

Create sector feature to categorize the employers where suspects work

Create age_group feature to categorize the suspects’ age

Create frequency feature to capture total number of cases of each state

Create date, month, year features

Features for visualization:

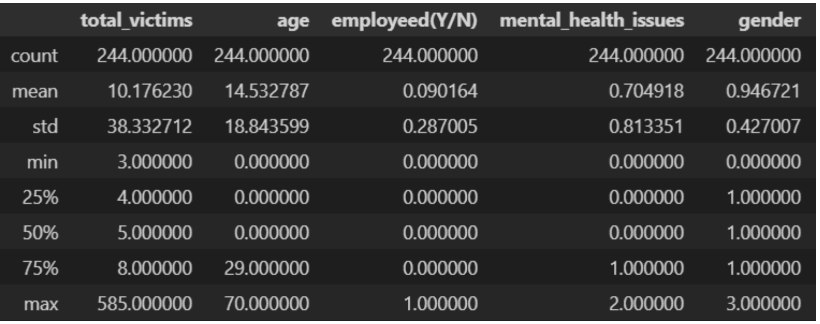
'total_victims', 'age', 'employeed(Y/N)', 'mental_health_issues', 'gender', 'licensed_business_entities', 'total_gun_deaths', 'state_population', 'unemployment_rate', 'sectors'

Project US Mass Shooting Crime Prediction

Exploration Data Analysis

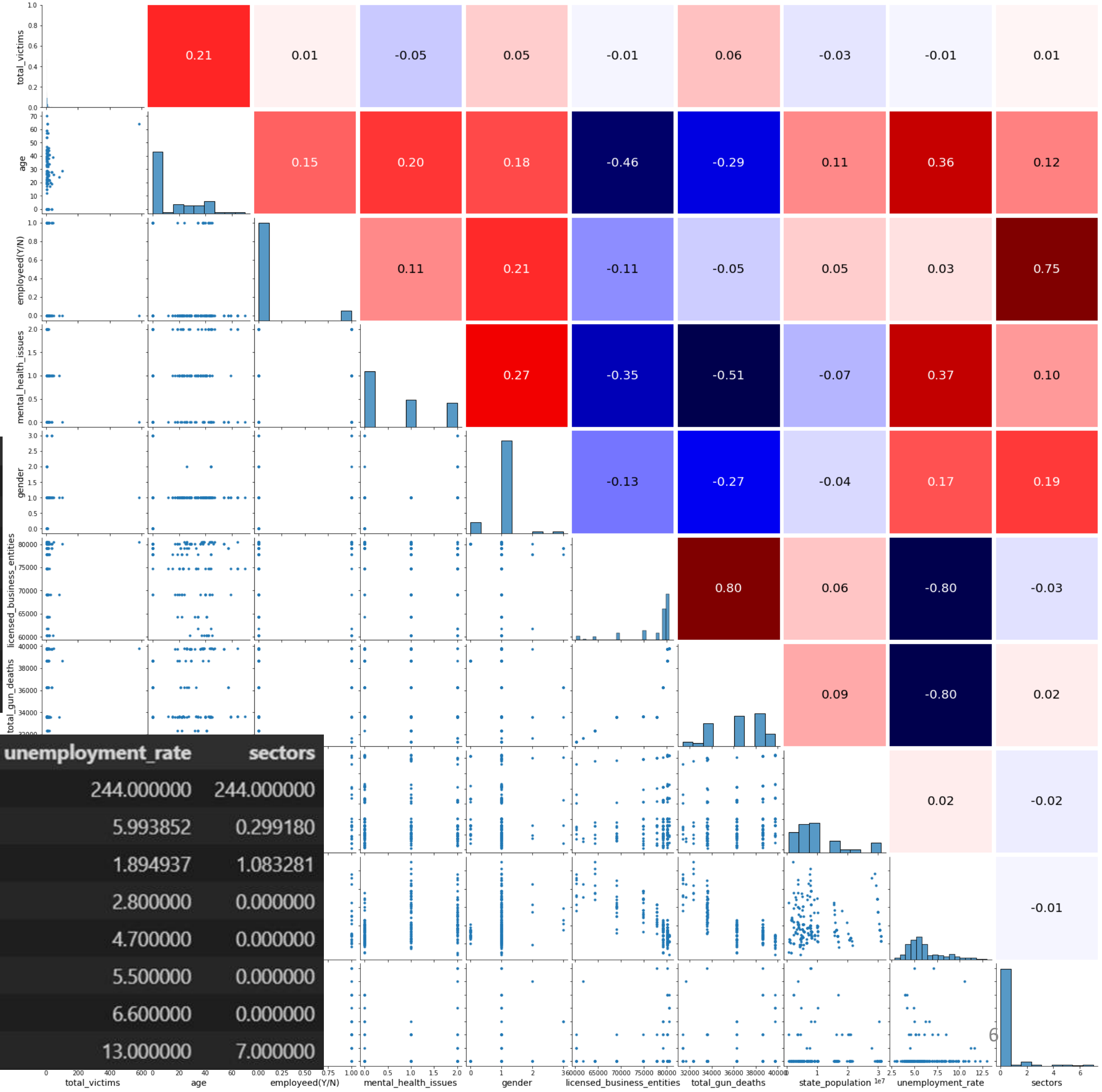
Target feature(total_victims) shows moderate to low correlation values to other predictors.

There are several fair to high correlation pairs among predictors: unemployment_rate vs total_gun_death & licensed_business_entities.



	total_victims	age	employeed(Y/N)	mental_health_issues	gender
count	244.000000	244.000000	244.000000	244.000000	244.000000
mean	10.176230	14.532787	0.090164	0.704918	0.946721
std	38.332712	18.843599	0.287005	0.813351	0.427007
min	3.000000	0.000000	0.000000	0.000000	0.000000
25%	4.000000	0.000000	0.000000	0.000000	1.000000
50%	5.000000	0.000000	0.000000	0.000000	1.000000
75%	8.000000	29.000000	0.000000	1.000000	1.000000
max	585.000000	70.000000	1.000000	2.000000	3.000000

	licensed_business_entities	total_gun_deaths	state_population	unemployment_rate	sectors
count	244.000000	244.000000	2.440000e+02	244.000000	244.000000
mean	77285.282787	36521.954918	1.012367e+07	5.993852	0.299180
std	5080.585517	2472.757342	8.174163e+06	1.894937	1.083281
min	60349.000000	31347.000000	5.380540e+05	2.800000	0.000000
25%	77815.000000	33636.000000	4.727283e+06	4.700000	0.000000
50%	79188.000000	36252.000000	7.758751e+06	5.500000	0.000000
75%	80119.000000	38658.000000	1.564546e+07	6.600000	0.000000
max	80493.000000	39773.000000	3.098267e+07	13.000000	7.000000

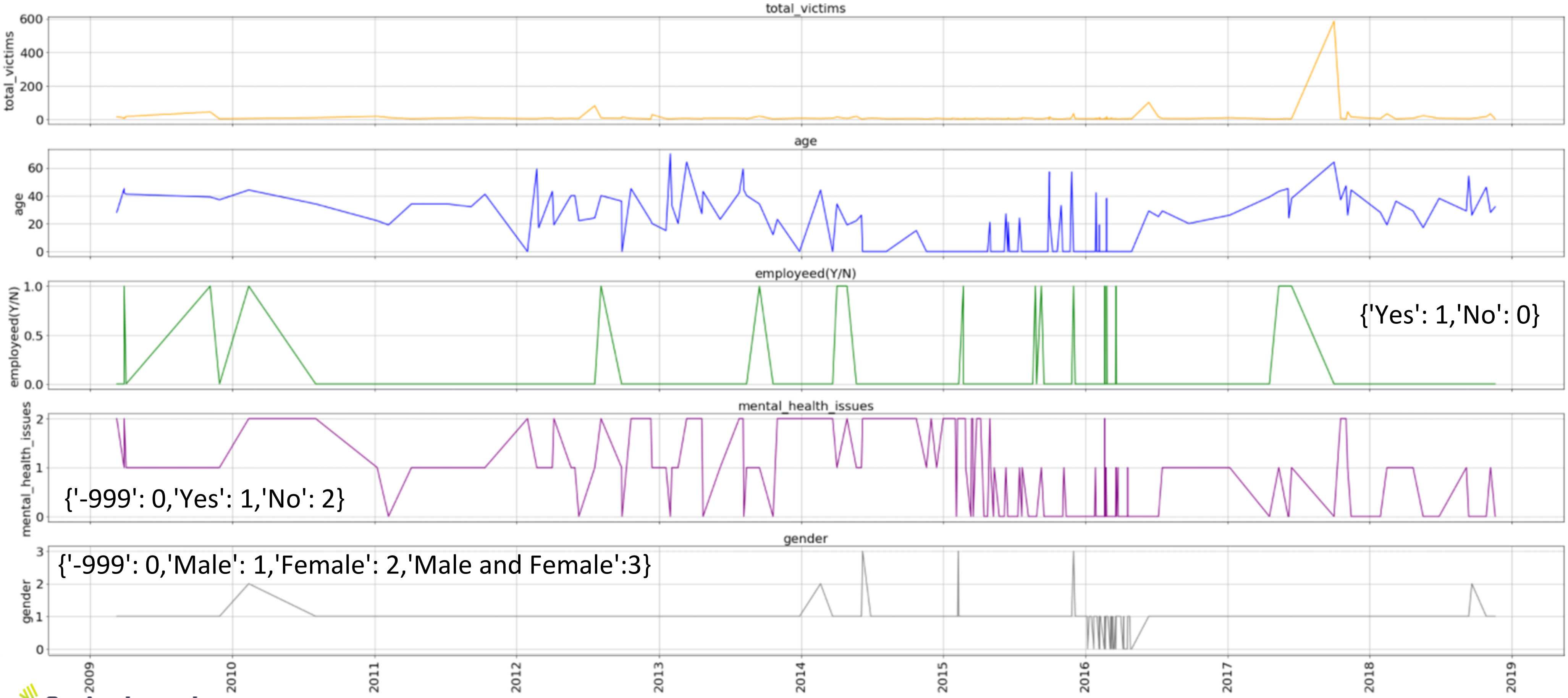


Project US Mass Shooting Crime Prediction

Exploration Data Analysis

Features visualization by time: 2009-2018:

- Number of victims significantly raised up in end of 2017
- Suspect's dominant age range is 15-45
- Most of shooters are male & jobless



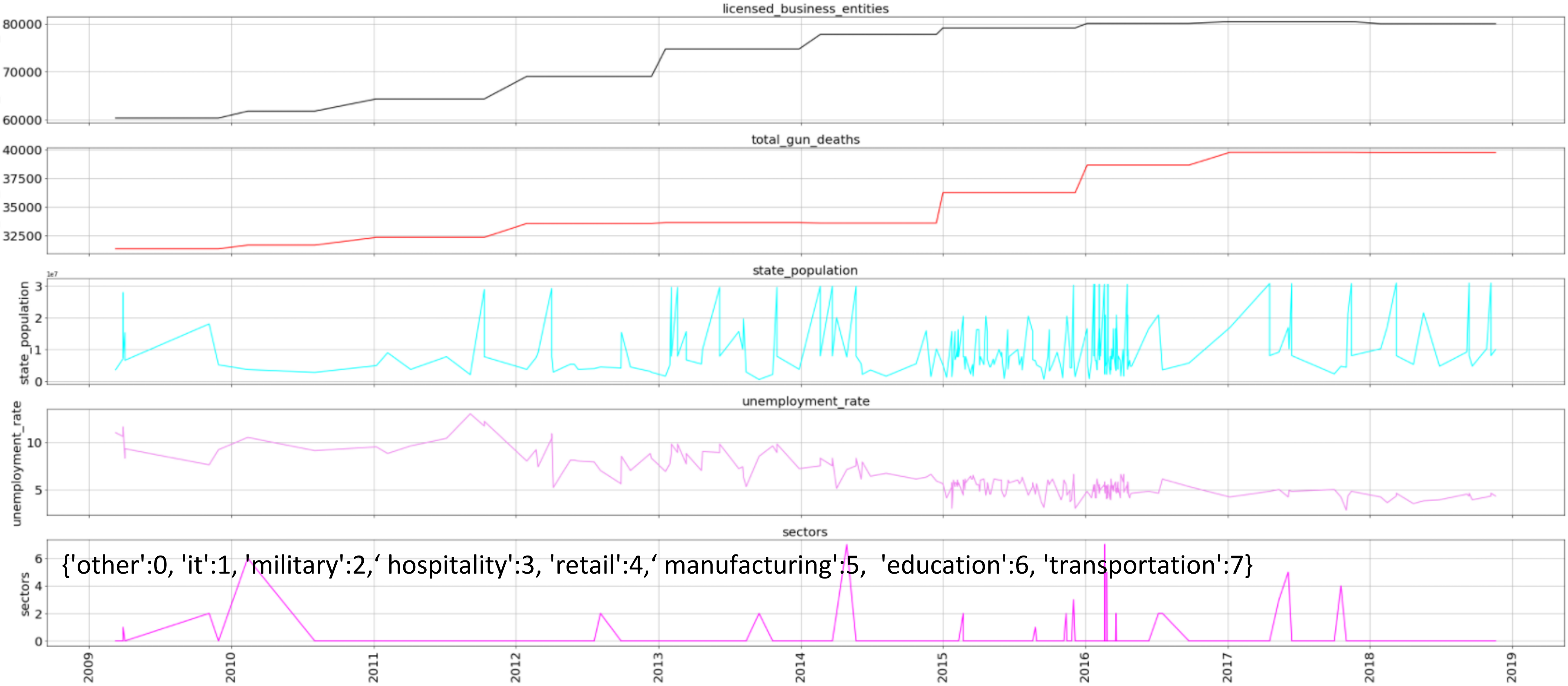
Project US Mass Shooting Crime Prediction

Exploration Data Analysis

Features visualization by time: 2009-2018:

- There is a strong positive relationship between number of gun licenses & 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.



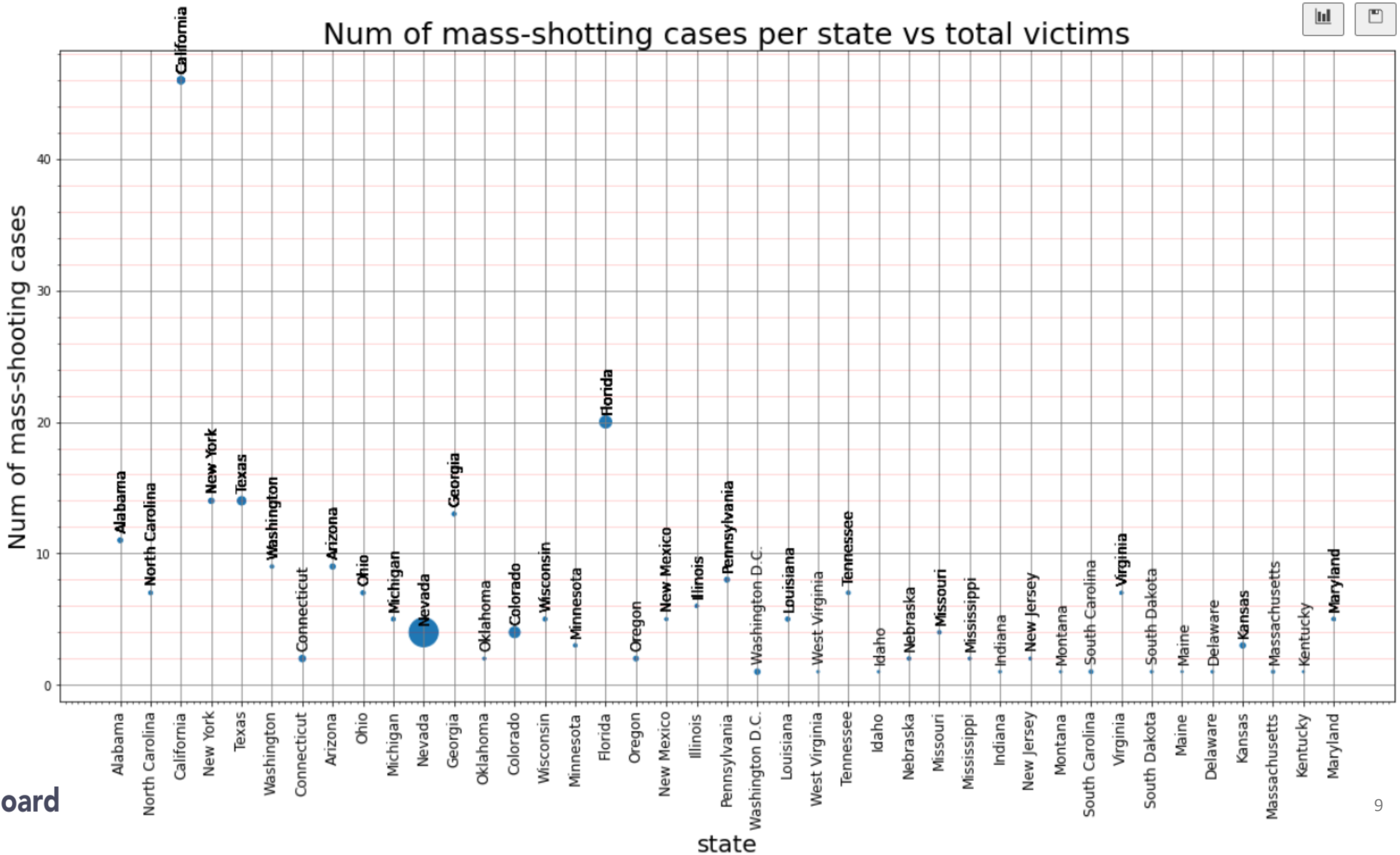
{'other':0, 'it':1, 'military':2,‘ hospitality':3, 'retail':4,‘ manufacturing':5, 'education':6, 'transportation':7}

Project US Mass Shooting Crime Prediction

Exploration Data Analysis

Features visualization by total_victims, states & frequency: 2009-2018:

- 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 are Florida & Colorado



Project US Mass Shooting Crime Prediction

Feature Selection

Data & features

Data shape (244x58), with dummy features from state (mother feature), has been divided to predictors(X) & target(y) = total_victims.

They are then has been spitted to 70% for training set & 30% for testing set. Their shapes X_train, X_test, y_train, y_test are respectively: (170, 57) (74, 57) (170,) (74,)

Such train & test sets are the inputs feeding to 5 different models in the next steps:

- Linear Regression
- Lasso Regression
- Ridge Regression
- OLS Regression
- RandomForest Regression

Project US Mass Shooting Crime Prediction

Modeling – Linear Regression

Key metrics:

R-square = 0.4978

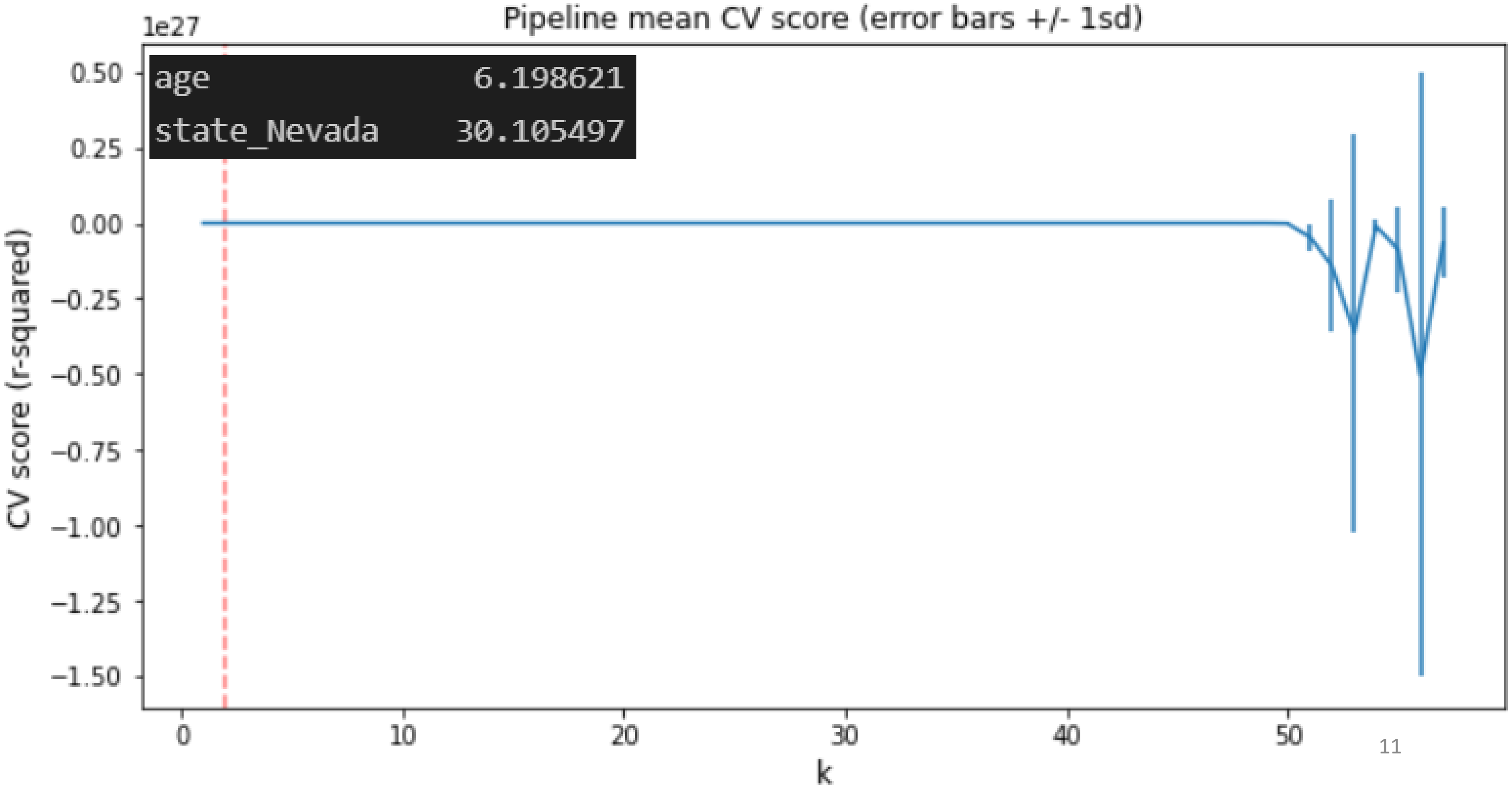
MAE = 14.7536

RMSE = 48.3944

Cross validation cv=5, metrics on test set

Mean_MAE = 11.9741, Mean_RMSE = 44.8658

Std_MAE = 6.9656, Std_RMSE = 44.2380



```
age              6.198621
state_Nevada    30.105497
```

Project US Mass Shooting Crime Prediction

Modeling – Linear Regression vs Lasso, Ridge

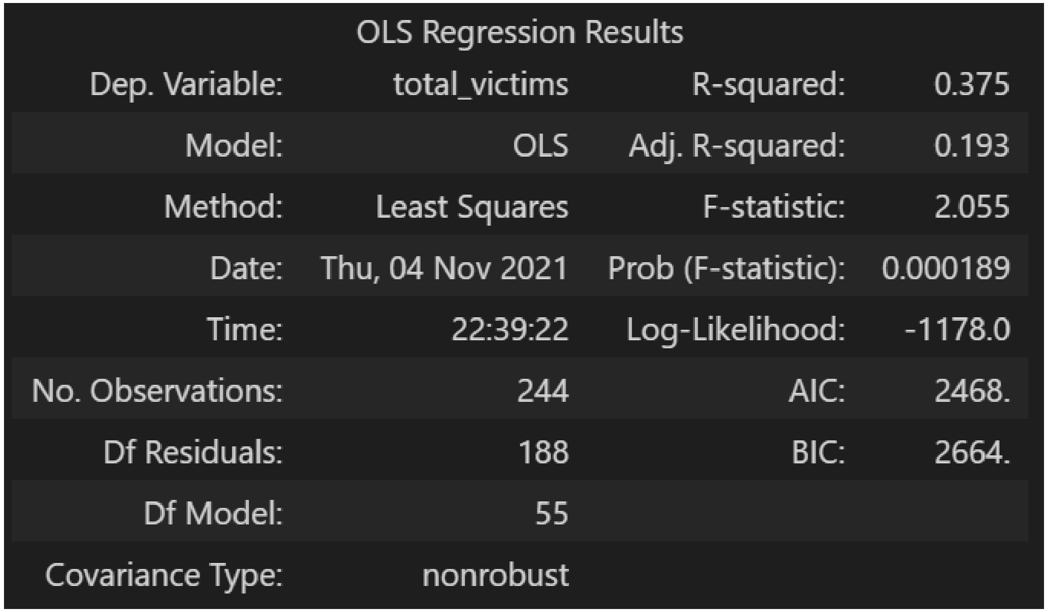
```
lasso_params = {'alpha':[0.02, 0.024, 0.025, 0.026, 0.03]}
ridge_params = {'alpha':[200, 230, 250,265, 270, 275, 290, 300, 500]}
```

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

Linear Regression & Lasso seem overfit, meanwhile Ridge seem underfit

Project US Mass Shooting Crime Prediction

Modeling – Ordinary Least Square Regression



OLS Regression Results

Dep. Variable:	total_victims	R-squared:	0.375
Model:	OLS	Adj. R-squared:	0.193
Method:	Least Squares	F-statistic:	2.055
Date:	Thu, 04 Nov 2021	Prob (F-statistic):	0.000189
Time:	22:39:22	Log-Likelihood:	-1178.0
No. Observations:	244	AIC:	2468.
Df Residuals:	188	BIC:	2664.
Df Model:	55		
Covariance Type:	nonrobust		

Key metrics

R2_train = 0.4926
R2_test = -2.0108
MAE_test = 14.7862
RMSE_train = 31.8737
RMSE_test = 26.0667

Statistic significant features (pvalue <5%)

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

Project US Mass Shooting Crime Prediction

Modeling – RandomForest Regression

Cross validation cv=5, metrics on test set

Mean_MAE = 10.8882, Mean_RMSE = 10.8882

Std_MAE = 6.1939, Std_MRSE = 34.4004

Deterministic metrics

R2_train = 0.7601

R2_test = -3.1089

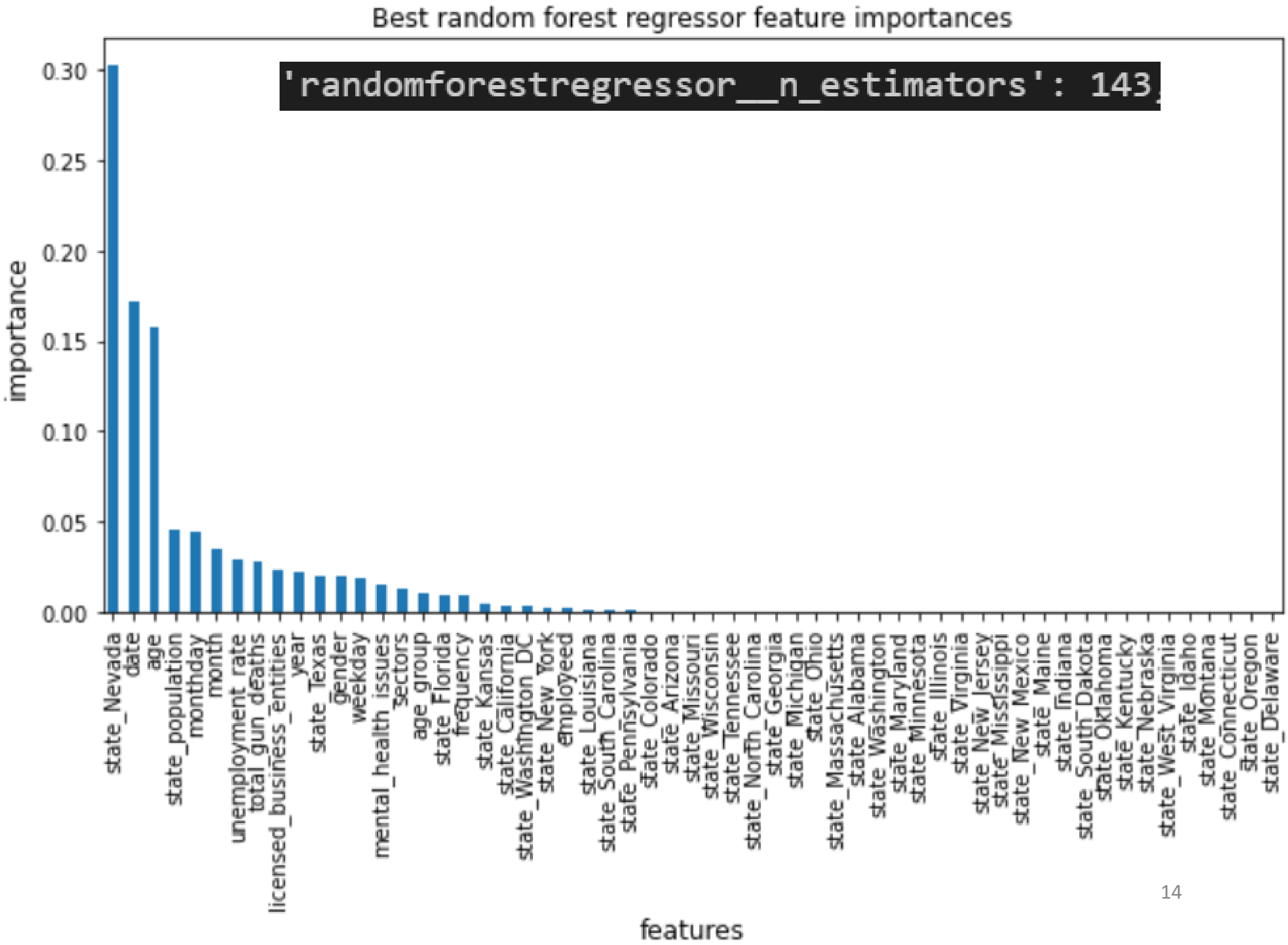
MAE_test = 10.0009

RMSE_train = 21.9143

RMSE_test = 30.4516

Most effected features:

state_Nevada, date, age, state_population, monthday, month, unemployment_rate, total_gun_death, license_business_entities


Best random forest regressor feature importances

```
'randomforestregressor__n_estimators': 143
```

Springboard

Project **US Mass Shooting Crime Prediction**

Results – Models Analysis

3 types of linear regression models have been conducted with 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. Number 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

Project US Mass Shooting Crime Prediction

Results – Model selection

Cross validation cv=5, metrics on test set

Mean_MAE = 9.7320 Mean_RMSE = 9.7312

Std_MAE = 6.3142 Std_MRSE = 35.9127

Deterministic metrics

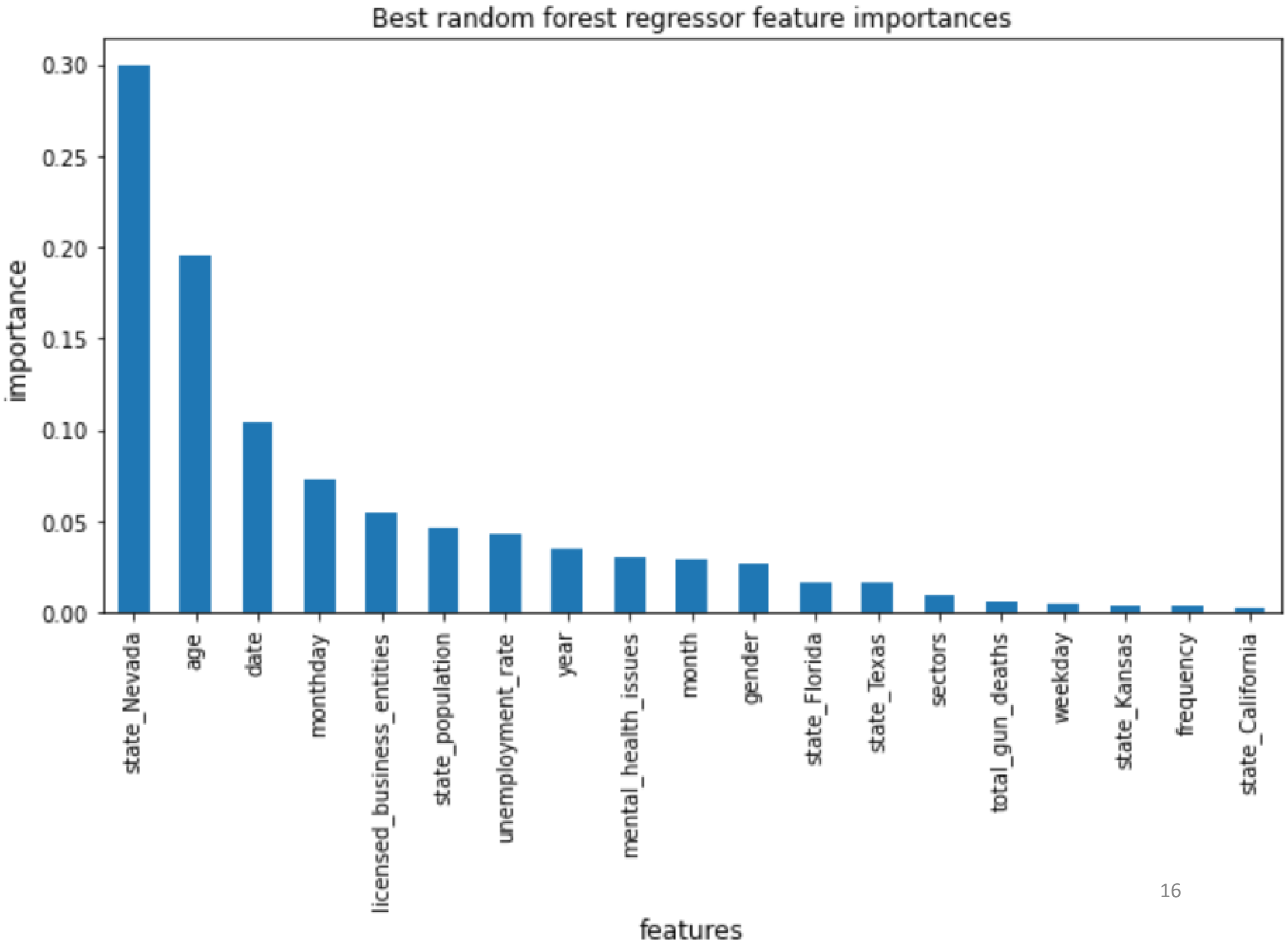
R2_train = 0.7717

R2_test = -3.2392

MAE_test = 10.0009

RMSE_train = 10.4332

RMSE_test = 21.3797

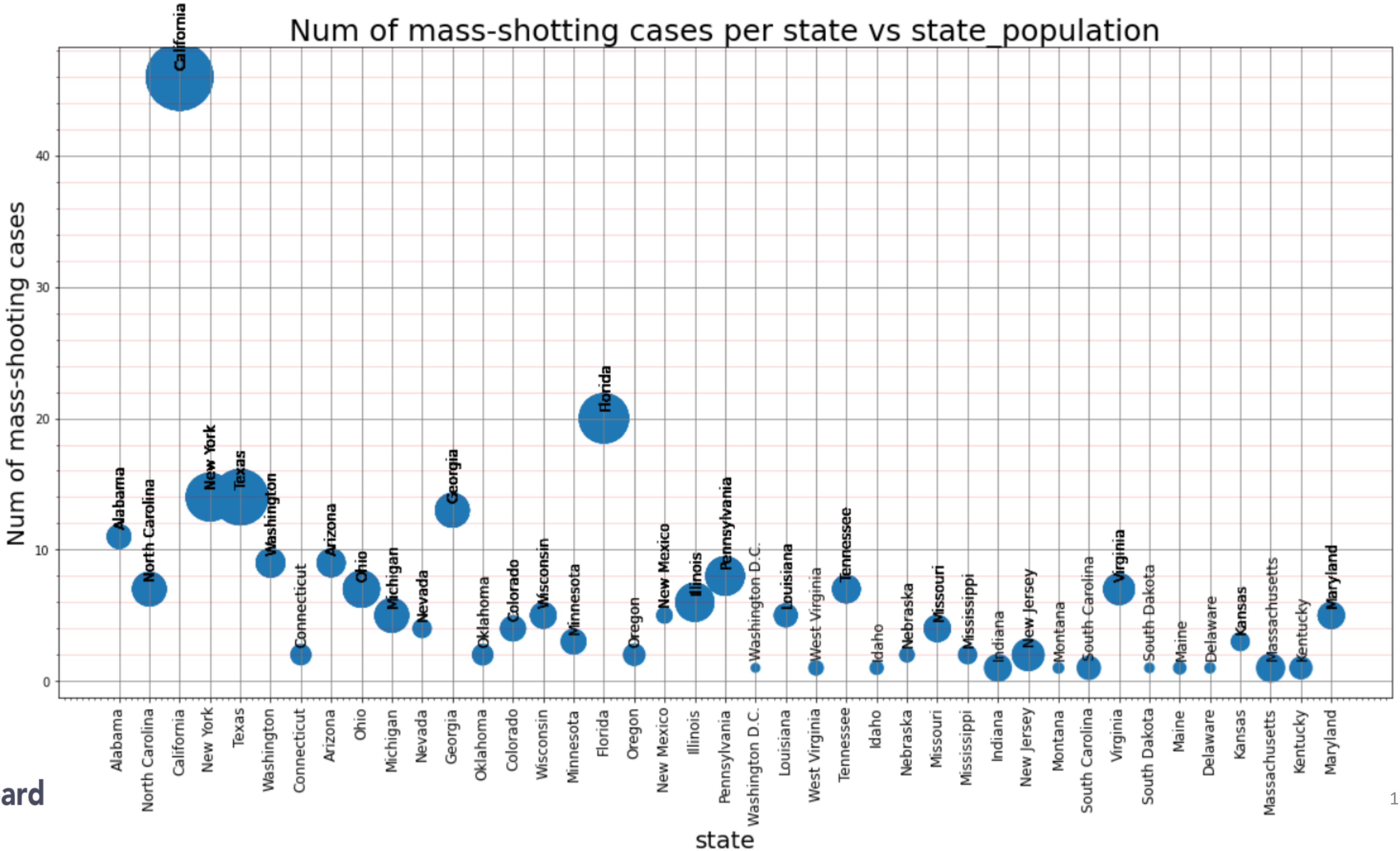


Springboard

Project US Mass Shooting Crime Prediction

Results – Features Explanation

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 number of cases .

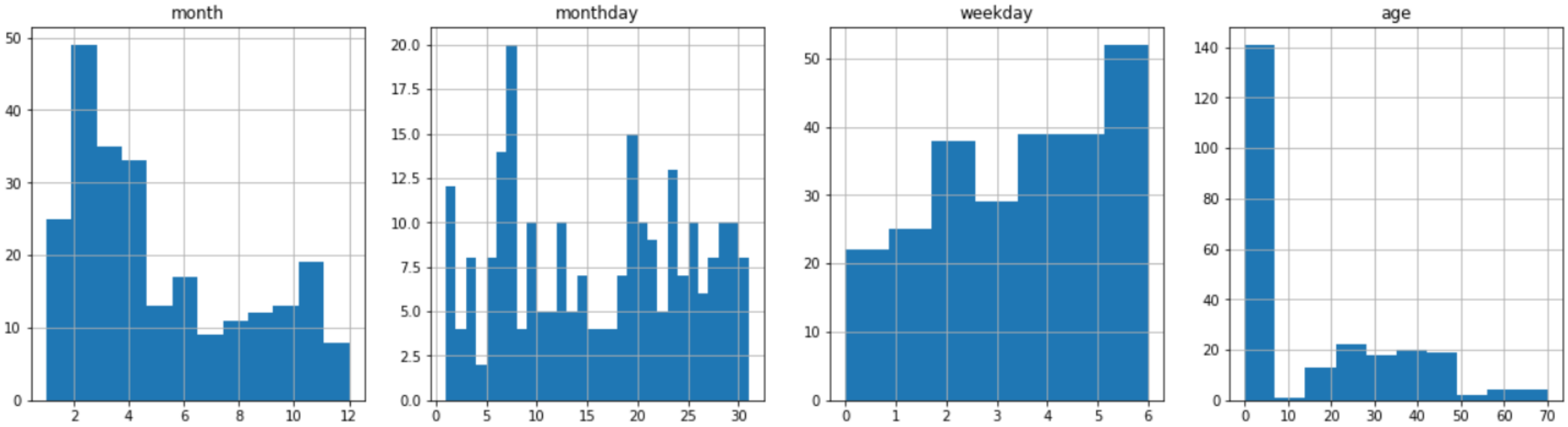


Springboard

Project US Mass Shooting Crime Prediction

Results – Features Explanation

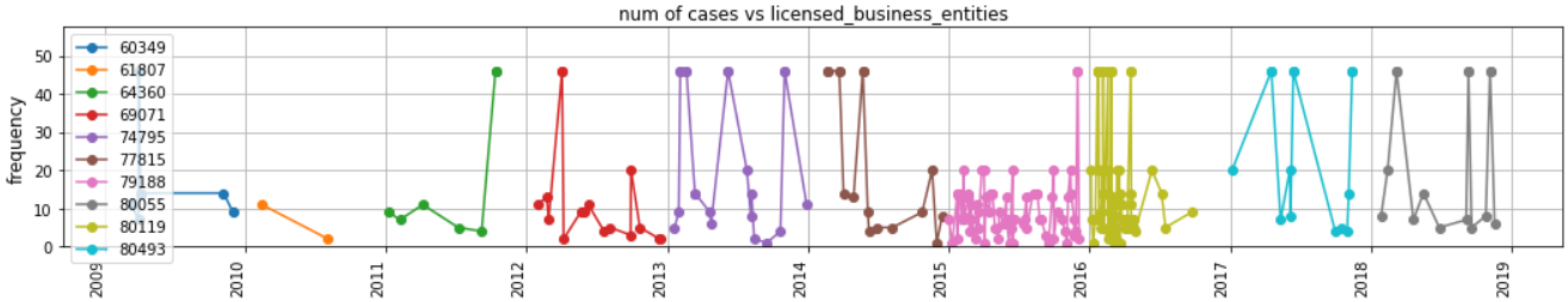
Mass shooting cases shows the high rate at the weekends & in Spring.
Shooter’s age range drops between 15-49. Age range from 0-5 could be an outlier



Project US Mass Shooting Crime Prediction

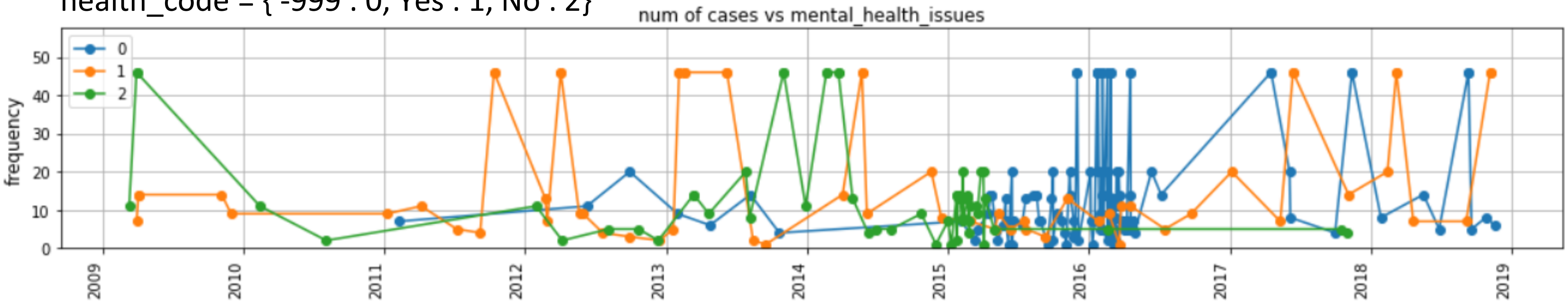
Results – Features Explanation

There is a positive relationship in number of gun licenses with number of mass shooting cases



The number of mass shooting cases related to mental health issues seems increasing in the more recent years

health_code = {'-999': 0,'Yes': 1,'No': 2}



Project US Mass Shooting Crime Prediction

Conclusion

The conclusion could be extracted:

- 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.

- These are indicators for government on improving the employment rate & strictly manage gun licenses, specially in high population states.

- Spend more effort on educating young people, promote more social connections & healthy activities, keep 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 higher value features and of course more data.

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

Project US Mass Shooting Crime Prediction

Recommendation

Need a further discussion about the performance of both Linear & RF regression:

- Key important features of both regressions is state_Nevada which could be an outlier, or need a special engineering treatment
- Negative R-square on test set score
- Fair high MAE & RMSE values
- Need more date to verify outliers & extract stronger features
- The way to improve model performance