Project Presentation

Machine Learning model for analyzing the fatal Police killings in the USA

by- Tarun Rajora

Problem Statement

In Brief….

The Washington Post is tracking more than a dozen details about each killing - including the race, age etc.

By using certain Machine Learning Classification techniques we've to analyze the chances of different races involved in a fatal encounter.

Preview of the Dataset

There are total 5 datasets :

1. First dataset containing the median household income in all US cities.
2. Second dataset containing the high school graduation rate for people over 25 in all US cities.
3. Third dataset containing the percentage of people below Poverty Line.
4. Fourth dataset regarding racial demographic in all US cities.
5. Fifth dataset containing details of individuals encountered by Police.

Approach

DATA PREPROCESSING

FEATURE ENGINEERING

MODEL TRAINING

MODEL EVALUATION

CONCLUSION

Data Pre-processing

- Features like age , income etc contains several missing values.
- To fill missing values in continuous features, we use ‘mean’ and in categorical features, we use ‘most frequent’ approach.
- Categorical variables are encoded using label Encoder and One-Hot Encoding.

Feature Engineering

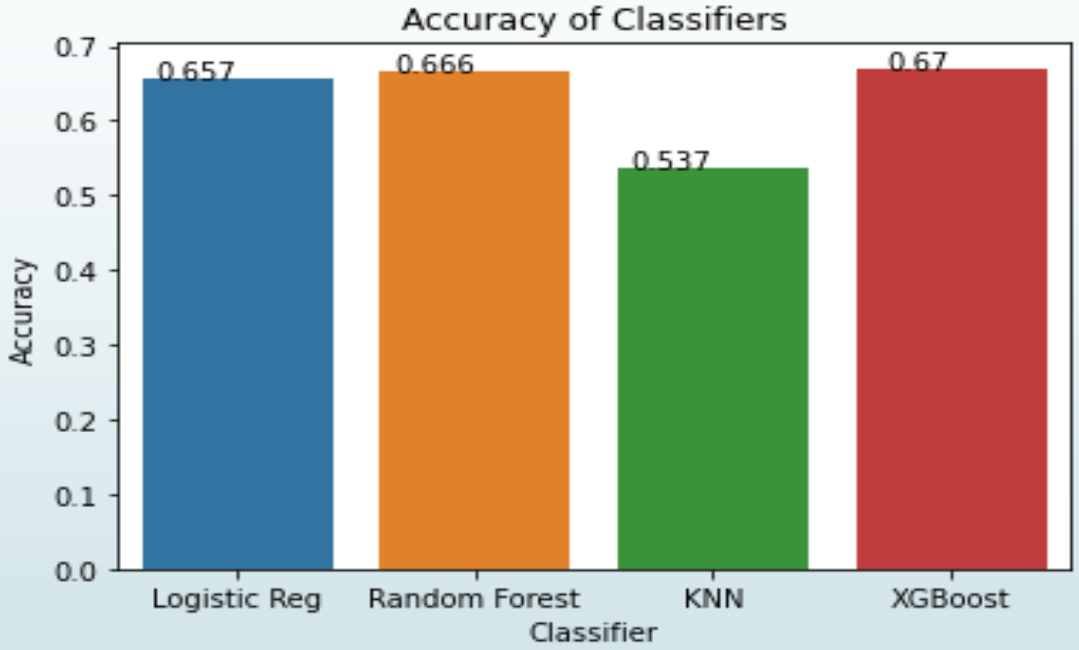
1. Here, first we remove the unwanted features such as name, id.
2. We merge two features: “city” and “state”/ “geographical area” common in all the 5 datasets into a single feature called “Place”.
3. To merge all the 5 datasets into a single one, we applied merge function from pandas library on the “Place” feature common to all.
4. After converting into a single dataset, we perform feature scaling.

Model Training

After analyzing the data it was clear that it is a classification problem since the output variable is a non-continuous type. So, the algorithms which we decided to work on are as follows:

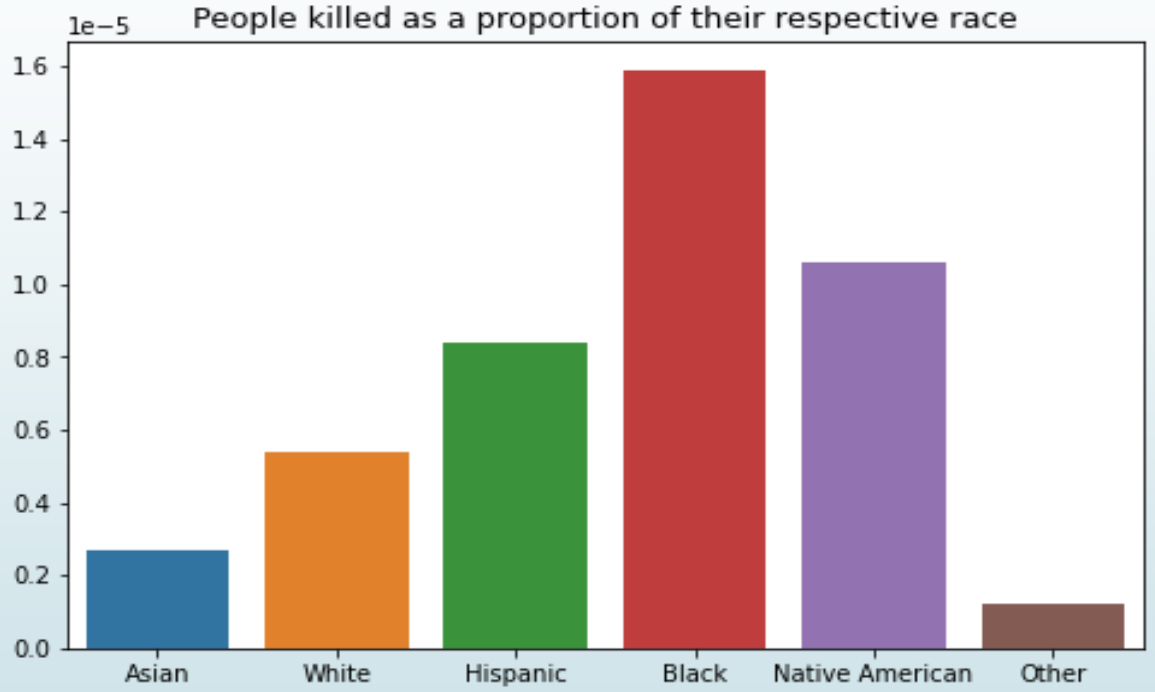
1. Logistic Regression
2. Random Forest Classifier
3. KNN algorithm
4. XGBoost Classifier

Model Evaluation



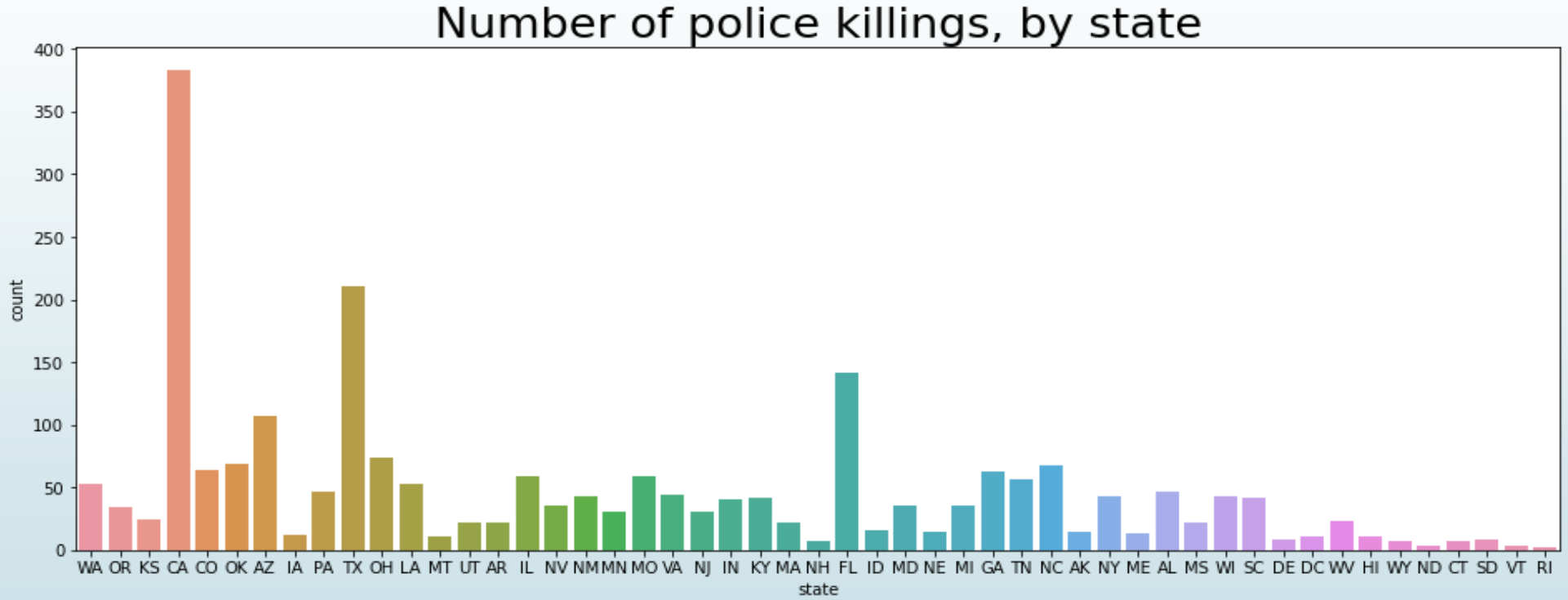
1. As we can see that the accuracies of Logistic regression, Random Forest and XGBoost are quite similar to each other. But XGBoost has a slightly higher accuracy than all.
2. Therefore, XGBoost is the best model for this dataset.

Conclusion



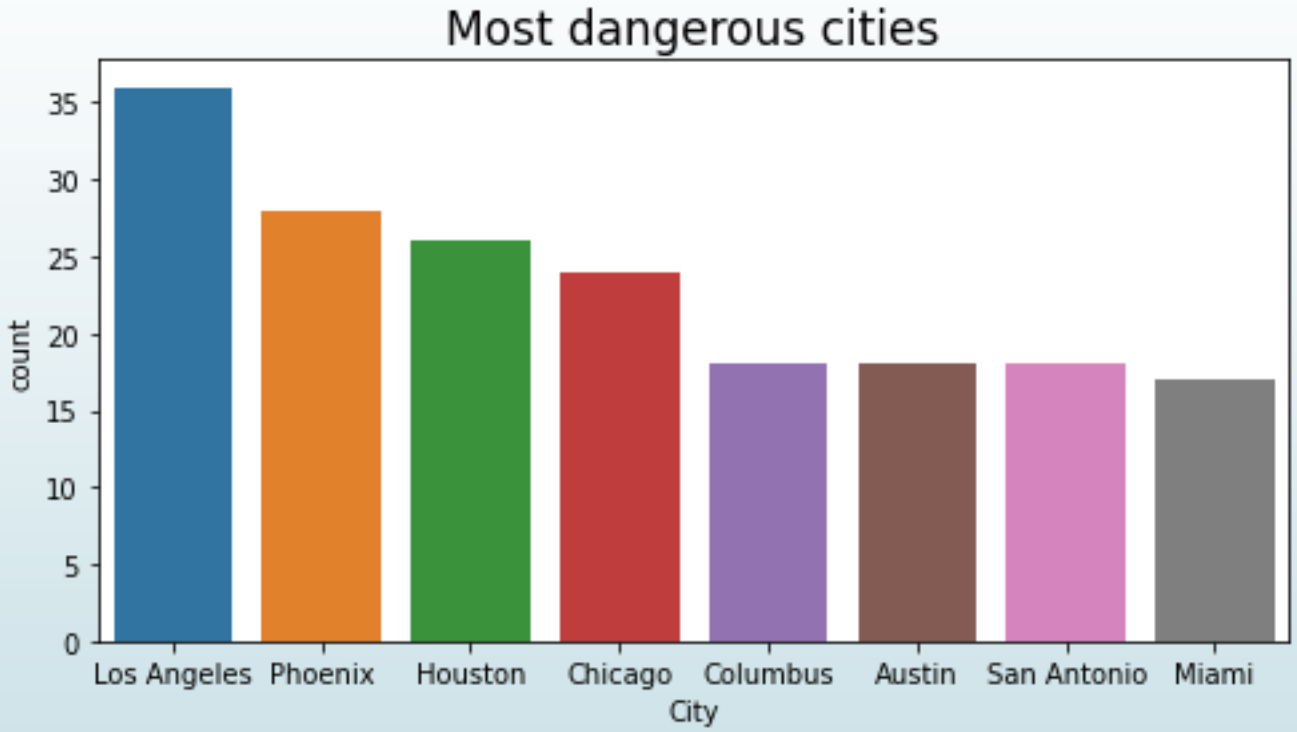
- This bar chart shows the number of victims per race as a proportion of the total US population of respective race.
- Blacks are 3 times more likely to become victims of police shootings than Whites.

Conclusion



- California is the state with the most fatal police shootings.

Conclusion



- Los Angeles has the highest number of encounters and hence it is the most dangerous city.

References

- https://www.kaggle.com/rabiayapicioglu/fatal-police-shootings-in-us-data-analysis
- https://python-graph-gallery.com/barplot/
- https://scikit-learn.org/stable/
- https://stackoverflow.com/
- https://towardsdatascience.com/machine-learning/

Thank you