Analyzing the Boston Crime Dataset

December 5th, 2019

Team Members

Alexander Moore

Reetayan Das

Abubeker Abdullahi

The Data

- Crime Dataset provided by Boston Police Department(BPD) documented initial details surrounding an incident to which an officer responded.This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred.

- **Source**
  - Kaggle (https://www.kaggle.com/AnalyzeBoston/crimes-in-boston)
- **Range**
  - June 14th, 2015 to September 3rd, 2018

DataSet Description

Columns

- A INCIDENT_NUMBER
- # OFFENSE_CODE
- A OFFENSE_CODE_GROUP
- A OFFENSE_DESCRIPTION
- A DISTRICT
- # REPORTING_AREA
- SHOOTING
- OCCURRED_ON_DATE
- # YEAR
- # MONTH
- A DAY_OF_WEEK
- # HOUR
- A UCR_PART
- A STREET
- Lat
- # Long
- A Location

Influences :

[1]Which Crime is mostly likely or could be occurred?

[2] Which Time is most crucial for a crime to be occurred ?

[3] What Areas of Boston Could be impacted?

[4] What are the factors which could affect the different type of Crime?

Data-PreProcessing-1

Filter

	Var1	Freq
1	Aggravated Assault	7807
2	Aircraft	36
3	Arson	94
4	Assembly or Gathering Violations	955
5	Auto Theft	4851
6	Auto Theft Recovery	1051
7	Ballistics	981
8	Biological Threat	2
9	Bomb Hoax	75
10	Burglary - No Property Taken	2
11	Commercial Burglary	1337

```
sv = offense_table_order %>% filter(Freq>5000)
sv <- mutate(sv,violation_level = 'strong')

mv = offense_table_order %>% filter(1000<Freq & Freq<5000)
mv <- mutate(mv,violation_level = 'medium')

lv = offense_table_order %>% filter(500<Freq & Freq<1000)
lv <- mutate(lv,violation_level = 'low')

nv = offense_table_order %>% filter(Freq<500)
nv <- mutate(nv,violation_level = 'noviolation')

group_crime <- rbind(sv,mv,lv,nv)
offense_table_sorted <- merge.data.frame(group_crime,offense_table)
```

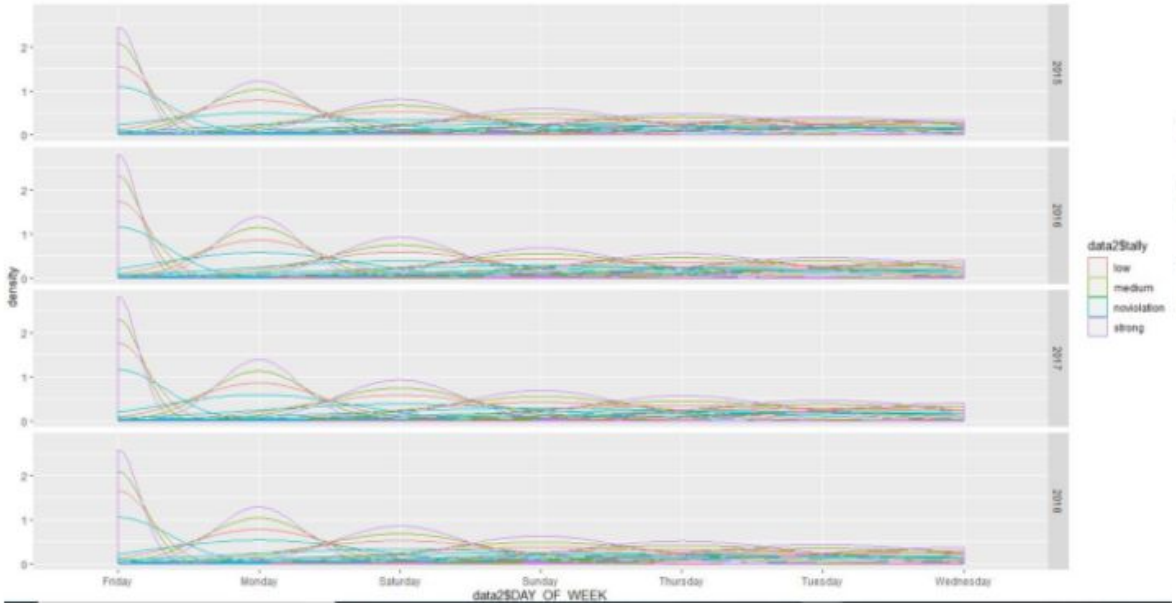
Data Preprocessing 2

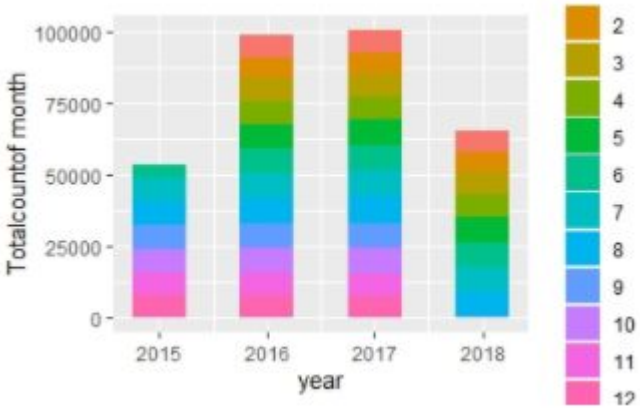
Var1	Freq	violation_level
Aggravated Assault	7807	strong
Aircraft	36	noviolation
Arson	94	noviolation
Assembly or Gathering Violations	955	low
Auto Theft Recovery	1051	medium
Auto Theft	4851	medium
Ballistics	981	low
Biological Threat	2	noviolation
Bomb Hoax	75	noviolation
Burglary - No Property Taken	2	noviolation
Commercial Burglary	1337	medium

STREET	Lat	Long	Location	tally
LINCOLN ST	42.35779	-71.13937	(42.35779134, -71.13937053)	strong
HECLA ST	42.30682	-71.06030	(42.30682138, -71.06030035)	strong
CAZENOVE ST	42.34659	-71.07243	(42.34658879, -71.07242943)	strong
NEWCOMB ST	42.33418	-71.07866	(42.33418175, -71.07866441)	strong
DELHI ST	42.27537	-71.09036	(42.27536542, -71.09036101)	strong
TALBOT AVE	42.29020	-71.07159	(42.29019621, -71.07159012)	strong
NORMANDY ST	42.30607	-71.08273	(42.30607218, -71.08273260)	medium
LAWN ST	42.32702	-71.10555	(42.32701648, -71.10555088)	strong
MASSACHUSETTS AVE	42.33152	-71.07085	(42.33152148, -71.07085307)	medium
LESLIE ST	42.29515	-71.05861	(42.29514664, -71.05860832)	strong
OCEAN VIEW DR	42.31958	-71.04033	(42.31957856, -71.04032766)	strong

Primary/Trend Analysis

1) Average Distribution of Crime over 4 Years on respective Months
2) Crime Intensity Graph over number of Days on 4 years

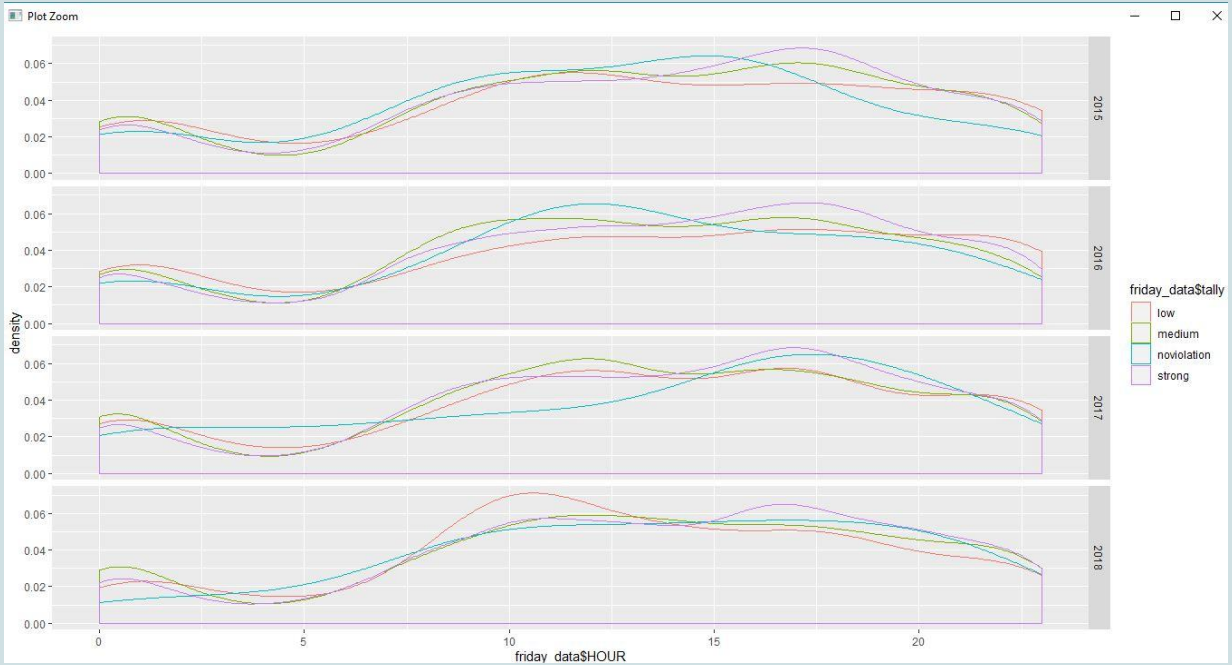




Friday has the highest occurrence of the strong violation crime

Analyses

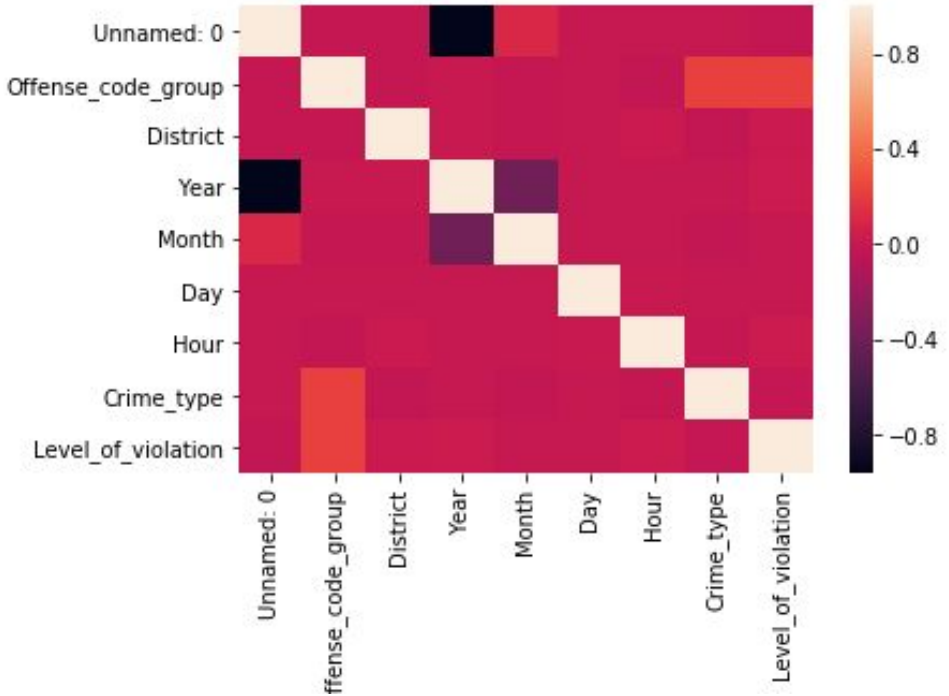
We have separated the Friday Data and wanted to see which time in the Friday has the most occurrence of strong level of violation here.
Graph shows that in 4 years every week of Friday in 17:30(Late afternoon) has the highest occurrence of strong violation



Selection Of predictors For the decision tree model

Correlation matrix shows that “Offense Code Group”, “District”, “Crime-type” has the high correlation value.

So we choose them as our predictors.



Prediction Model 1

Predictors : Crime_type, Offense_code_group
District ;Label = “Level_of_violation”

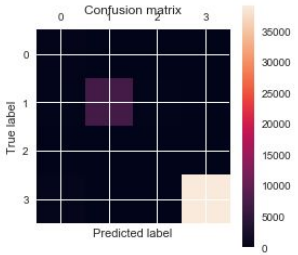
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

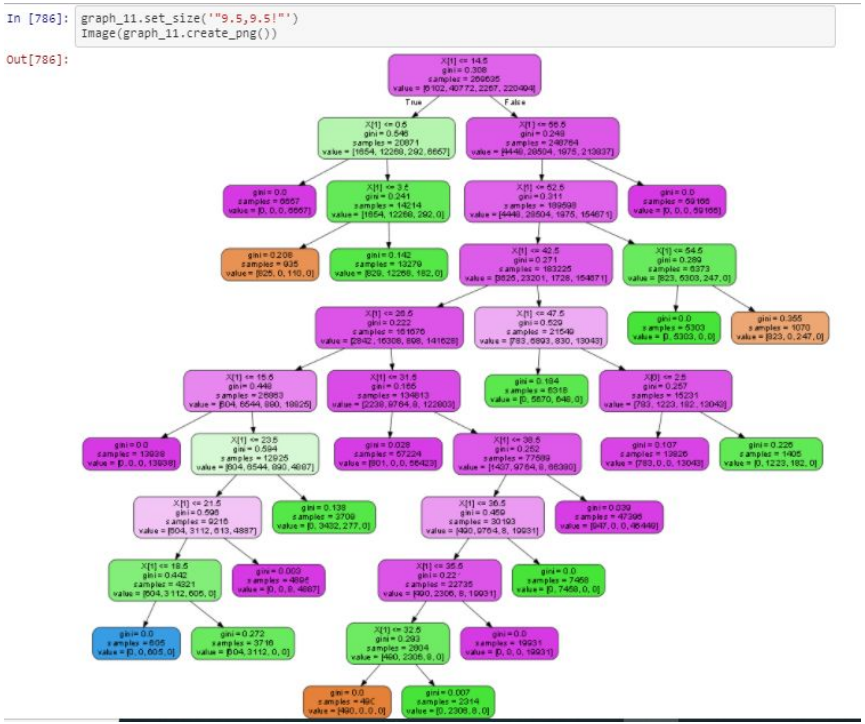
```
Accuracy: 0.9802240295903999
```

Note : District and Offense Code had to be encoded to 0 and 1 as they were not in numeric Types

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_pred,y_test)
print(cm)
plt.matshow(cm)
plt.title('Confusion matrix')
plt.colorbar()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```

```
[[  333     0    55     0]
 [  241  7109   187     0]
 [    0     0   124     0]
 [  458     0     0 39076]]
```





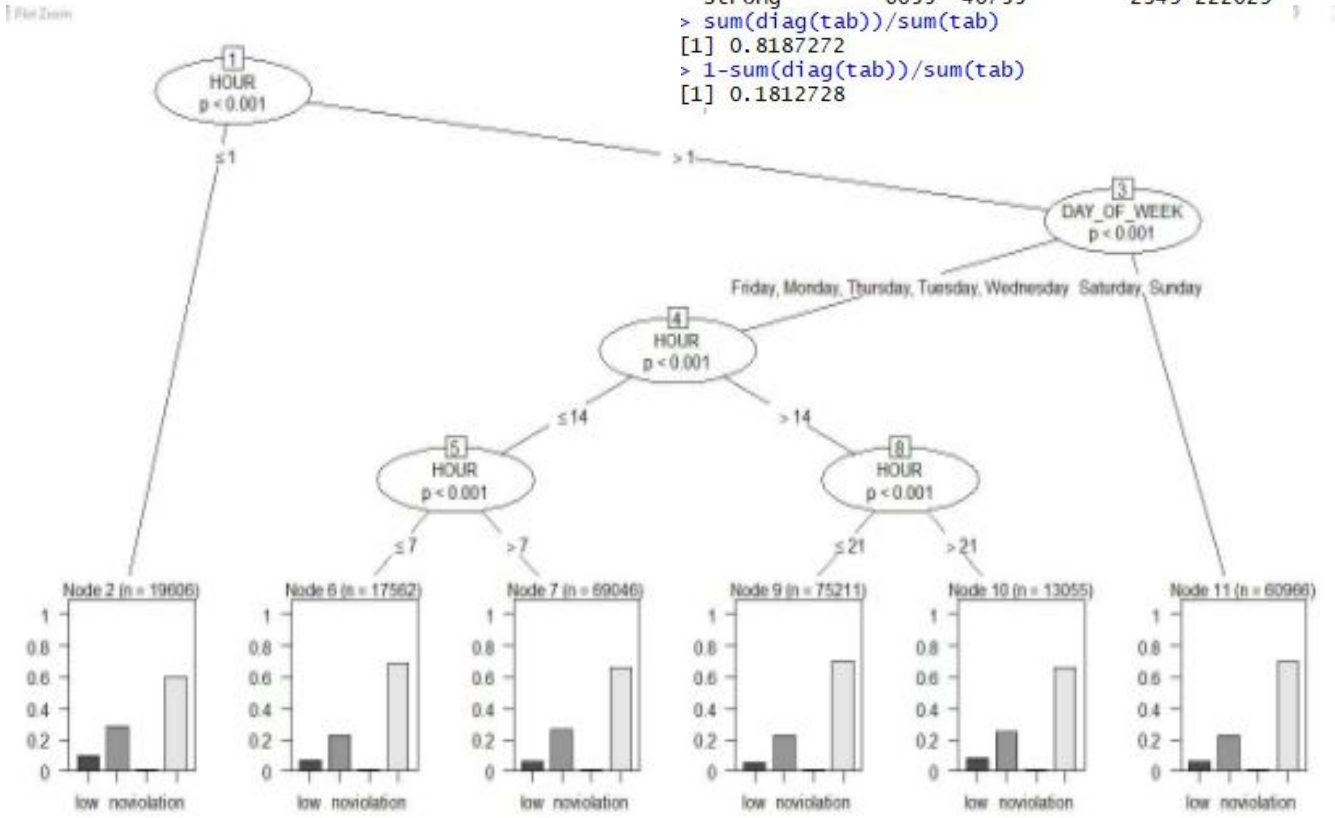
Separating WeekDays Vs Weekends Crime

```
> tab <- table(predict(tree),train2$tally)
> print(tab)

             low medium noviolation strong
  low          0      0           0      0
  medium       0      0           0      0
  noviolation  0      0           0      0
  strong    6055  40755        2349 222029
> sum(diag(tab))/sum(tab)
[1] 0.8187272
> 1-sum(diag(tab))/sum(tab)
[1] 0.1812728
```

This Decision tree tells about that a strong separation rule for weeknd crime and weekdays Crime.

Although Friday Data holds the most strong level violation it belongs to the weekdays Rule.



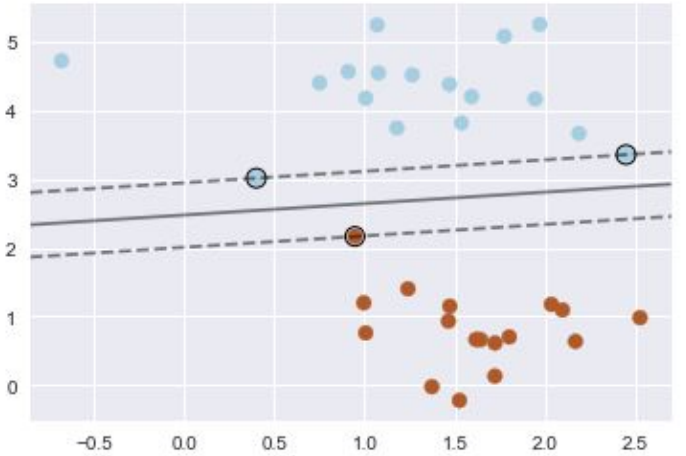
Prediction Model 2

Feature Clustering - District

We took the “District” Feature which have a high Feature importance and objective was to cluster them at least in two groups based on the level of violation(label = [1 to 5])

Applied Support Vector machine to find the Widest street between two clusters

```
<matplotlib.collections.PathCollection at 0x1f23f39ec88>
```



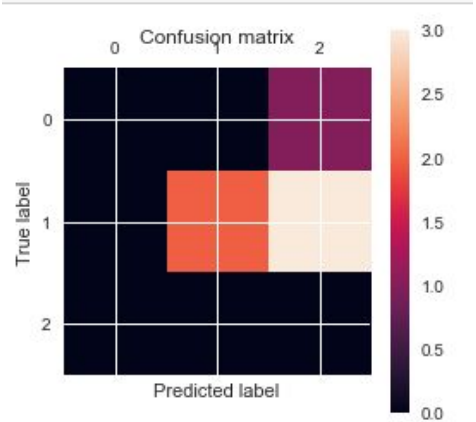

```
model.support_vectors_
```

```
array([[2.45161058, 3.35844964],
       [0.39920934, 3.01626962],
       [0.9461919 , 2.16566767]])
```

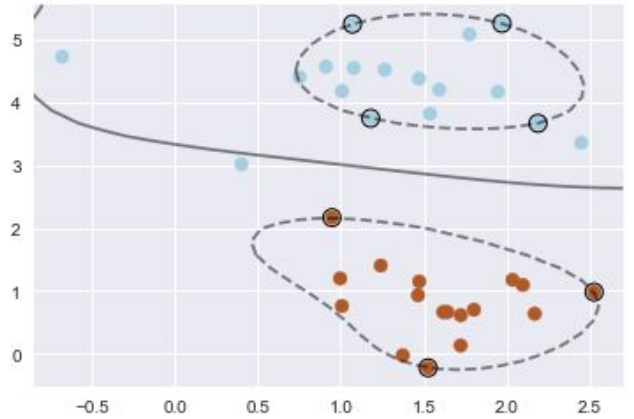
Feature Clustering - Offense codes

Linearity didnt exist when we introduced the “Offense code” as Feature.

Applied RBF kernel to plot the Decision Boundary here.




```
<matplotlib.collections.PathCollection at 0x1f23f9d3710>
```




```
model.support_vectors_
```

```
array([[ 2.19018277,  3.66855671],
       [ 1.17976408,  3.7486251 ],
       [ 1.0698984 ,  5.24906511],
       [ 1.97257657,  5.25887053],
       [ 1.52672244, -0.22442003],
       [ 0.9461919 ,  2.16566767],
       [ 2.52917639,  0.98150255]])
```

Analyses - Gradient Boosted Decision Tree Model

XGBoost model with a decision tree as its base learner was created to maximize prediction accuracy.

- R^2 = 91.6%

Using this model, feature importance was assessed

- Spatial features appeared more important than temporal features

Analysis with Census Data

Location Data was used to make requests for corresponding tracts with the Census API

- Rounded Coordinates to 3 places to reduce number of requests
- Census Tracts for coordinates recorded, joined with Crime Report dataset
- SES Data from 2015 ACS merged by Tract Number

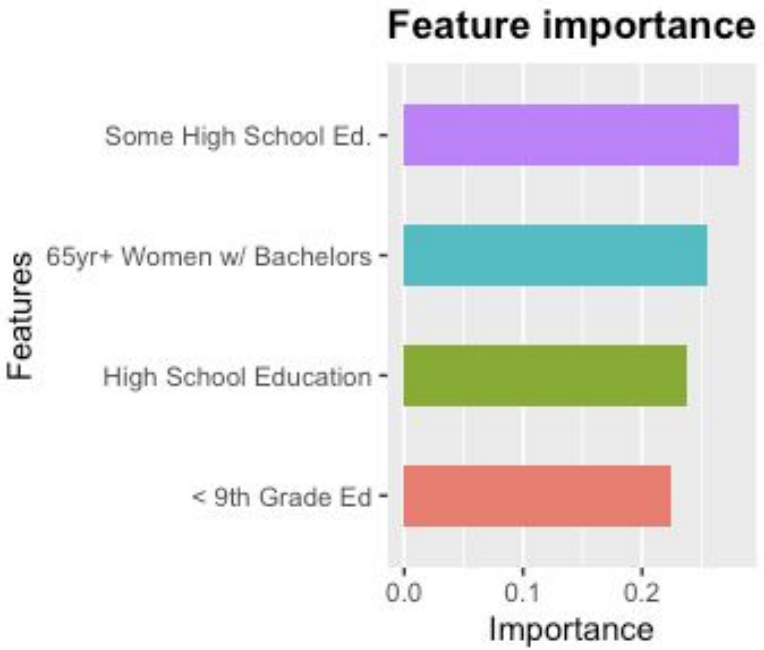
Analysis with Census Data

Tract level compositional differences for analyses -

- Age, Gender, Race, and Educational Attainment
- Model Accuracy - R^2 = 91.07%

The only

-



Conclusion