Seattle Crime

By Alejandro Harrison

Business Goal

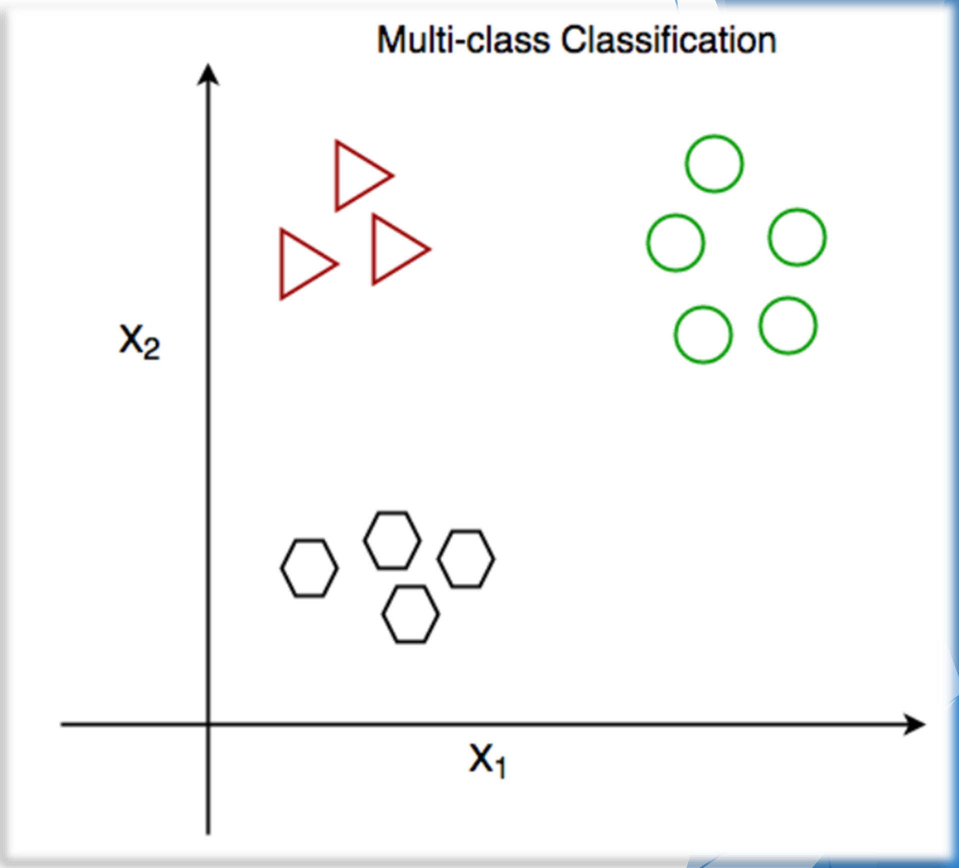
- Seattle large, growing city
- Major problem is crime
- As mayor, want to reduce crime to make city safer
- Need to be able to predict crime to reduce

Method

- Use of Multiclass Classification models
- Looking at past crimes
- Able to classify types and location of new crimes
- What is Multiclass Classification?

Multiclass Classification

- Dividing data into more than 2 classes/categories
- Can be used to predict future values
- Crime type and location



The Data

- Seattle.gov SPD crime dataset(2008-present)
- Over 1 million rows
- 17 variables such as crime date, crime category, crime location, etc
- Data categorical

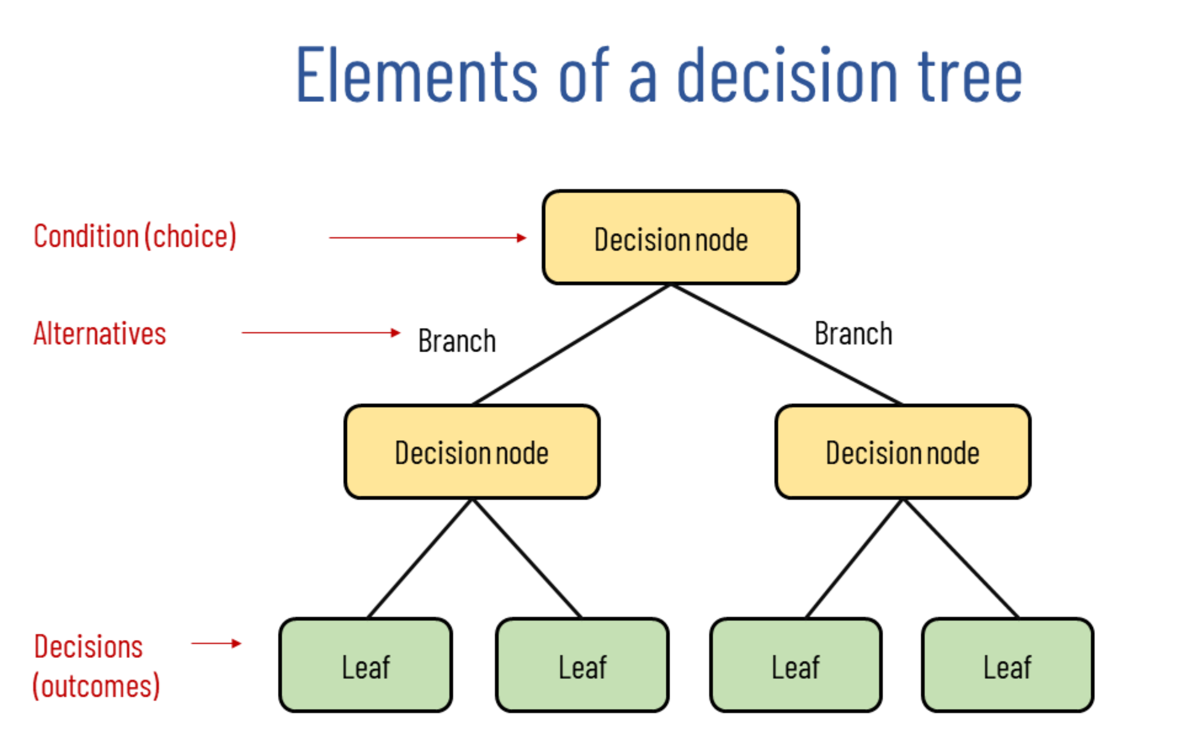
```
RangeIndex: 1036236 entries, 0 to 1036235
Data columns (total 17 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   Report Number           1036236 non-null  object
 1   Offense ID              1036236 non-null  int64
 2   Offense Start DateTime  1034989 non-null  object
 3   Offense End DateTime    583385 non-null   object
 4   Report DateTime         1036236 non-null  object
 5   Group A B               1036236 non-null  object
 6   Crime Against Category  1036236 non-null  object
 7   Offense Parent Group    1036236 non-null  object
 8   Offense                 1036236 non-null  object
 9   Offense Code            1036236 non-null  object
 10  Precinct                1036232 non-null  object
 11  Sector                  1036234 non-null  object
 12  Beat                    1036234 non-null  object
 13  MCPP                    1036236 non-null  object
 14  100 Block Address       992187 non-null   object
 15  Longitude               1036236 non-null  float64
 16  Latitude                1036236 non-null  float64
```

The Models

- Built two separate models
- Used two different algorithms
- CatBoost (gradient boosted decision tree)
- Random Forest

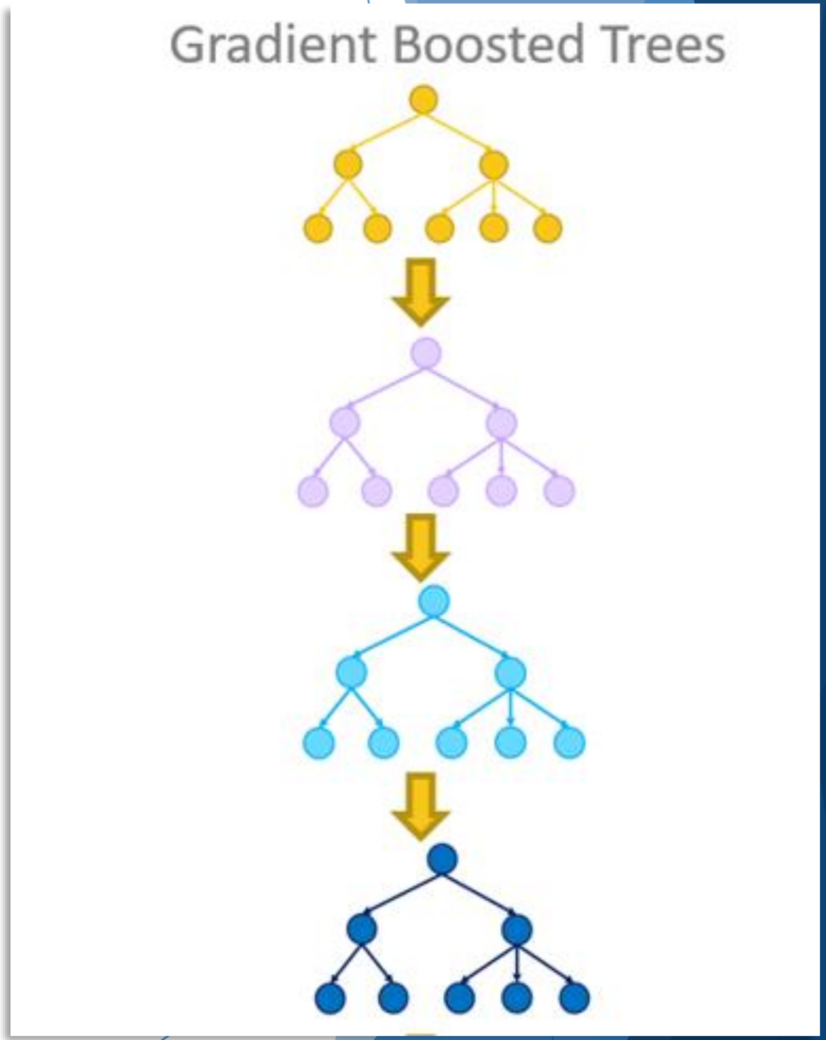
CatBoost

- Gradient boosted Decision Tree
- Used when data is categorical
- Decision tree splits data into series of decisions
- At end, final decision reached and data is categorized
- Decision Trees prone to overfitting



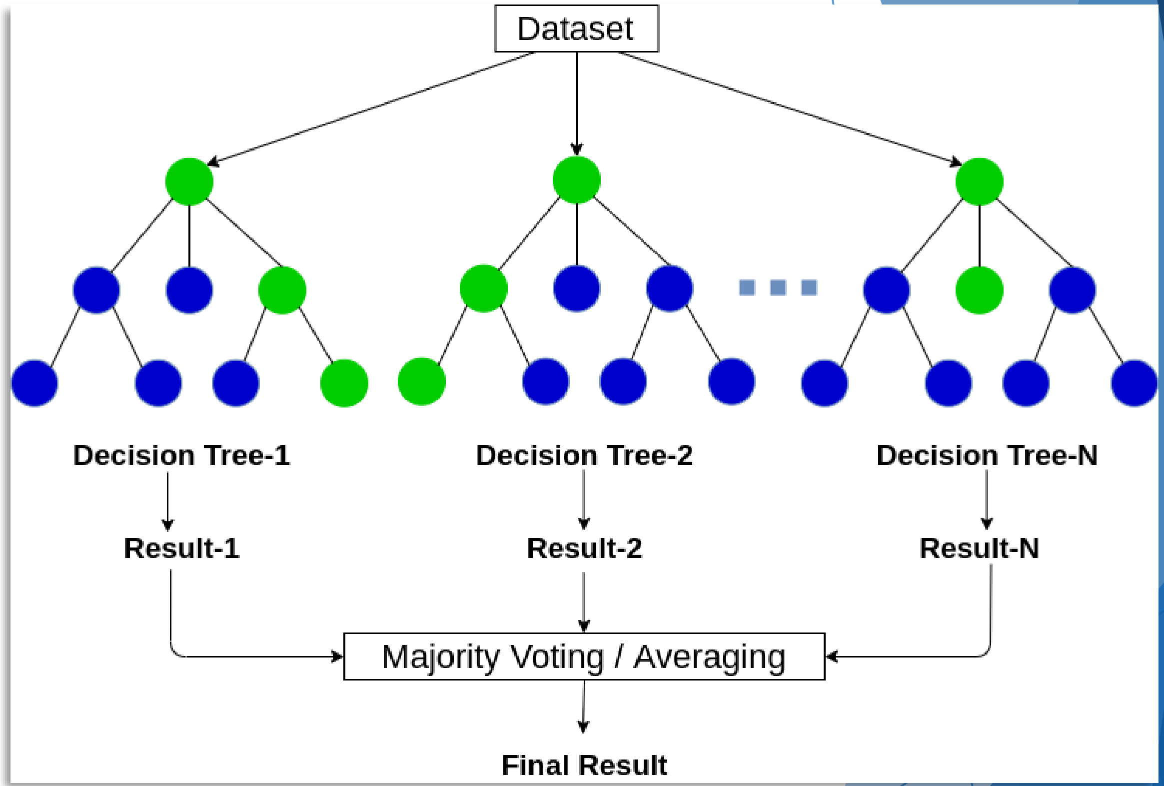
CatBoost Continued

- Reduce overfitting with gradient boosting
- Adding series of decision trees, each one improves upon last
- Other benefits include great default models and able to train on GPU



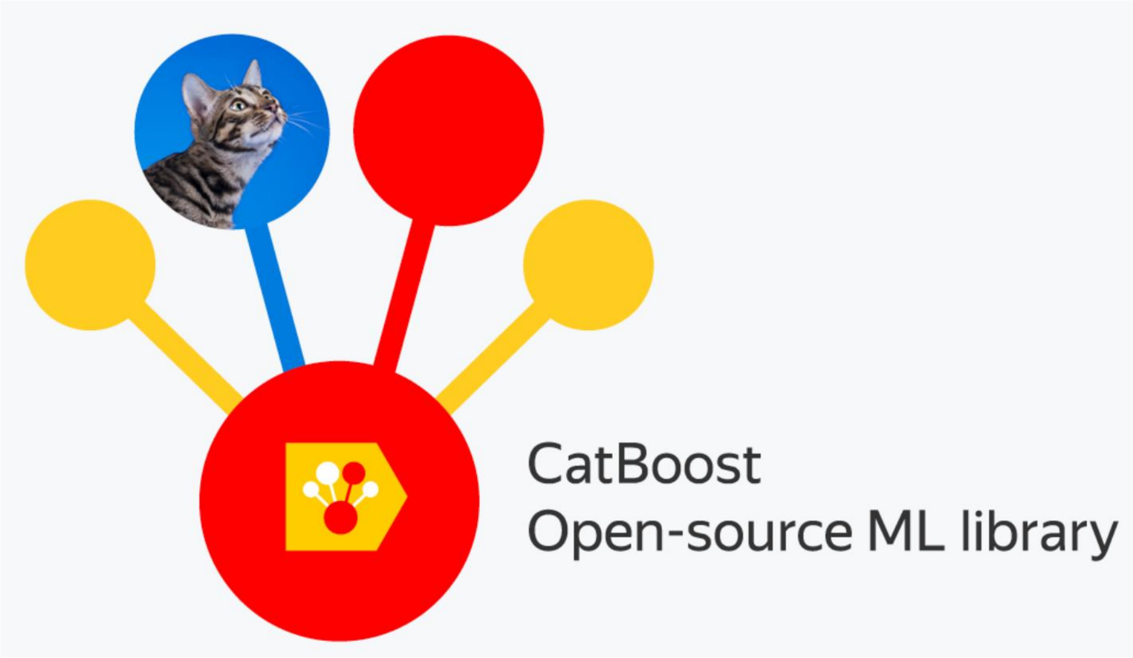
Random Forest

- Random samples created from data
- Individual decision trees made for each sample
- All tree predictions compared together
- Majority vote for best prediction



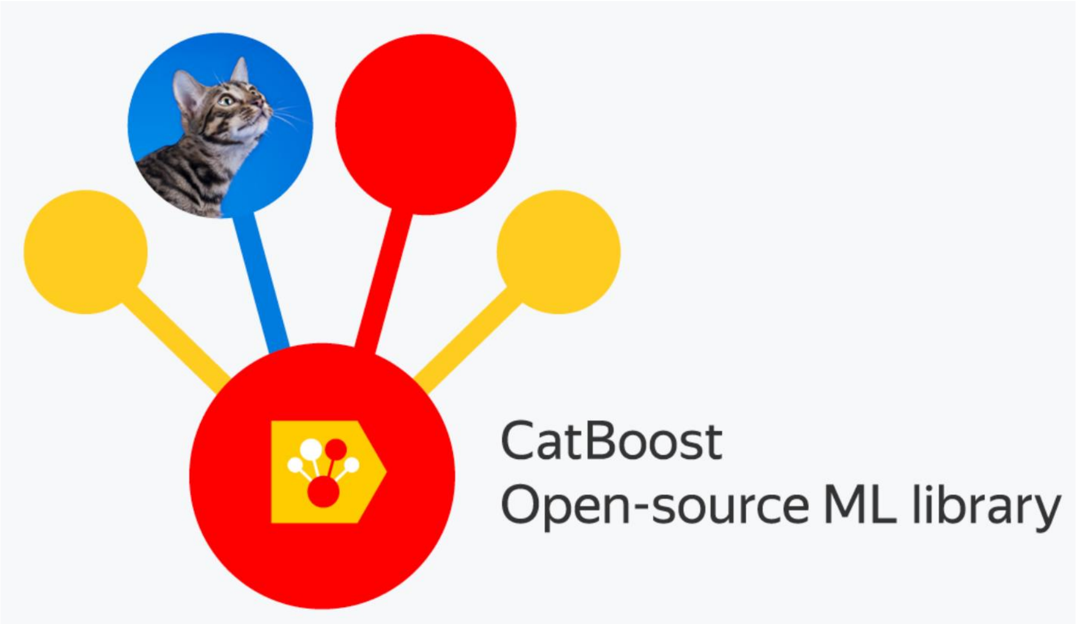
Crime Type Best Model

- Compared Random Forest and CatBoost models
- Found our CatBoost model to be the best model



Crime Location Best Model

- Compared a default CatBoost model and a tuned CatBoost model
- Found initial CatBoost model was our best one



Results

- Crime type, crime locations
- Recall, True Positives, False Negatives
- Total number of certain type crime/crime location, what percent did model correctly predict(Recall)
- How many crimes/crime locations were correctly predicted (TP)
- How many crimes/crime locations wrongly predicted as not belonging to a certain class (FN)

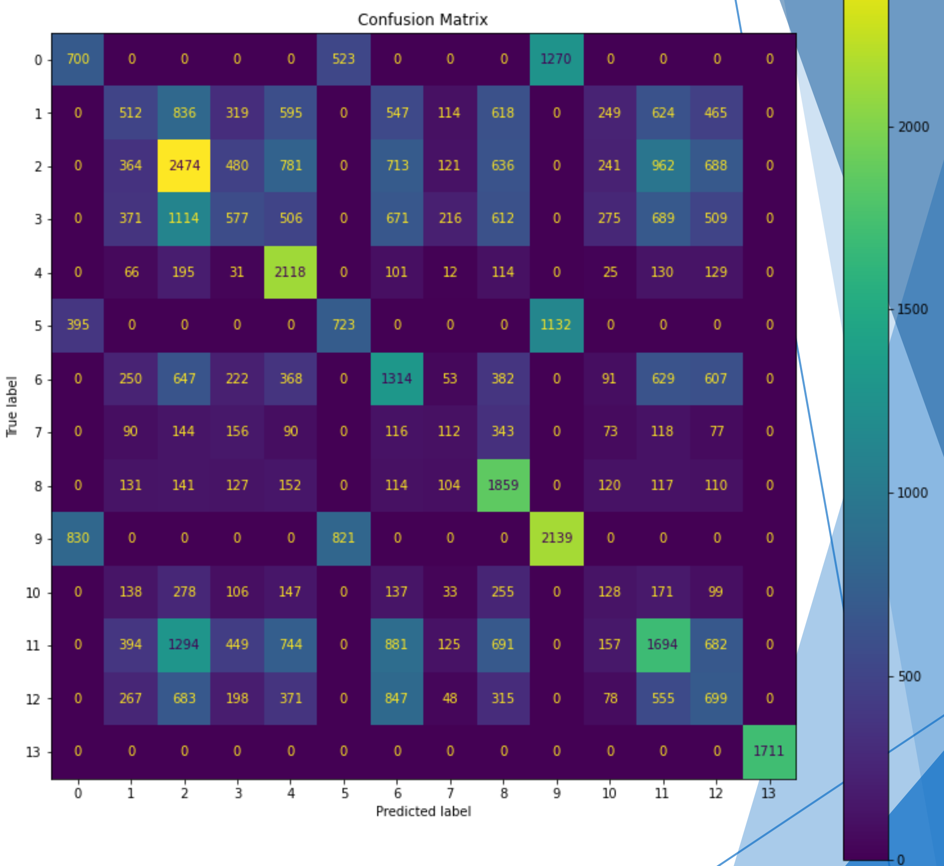
Crime Type Recall

- Best predicted crimes were Shoplifting and Simple Assault
- Shoplifting recall was .62
- Simple Assault recall was .56

	precision	recall	f1-score	support
Aggravated Assault	0.36	0.28	0.32	2493
All Other Larceny	0.20	0.10	0.14	4879
Burglary/Breaking & Entering	0.32	0.33	0.32	7460
Destruction/Damage/Vandalism of Property	0.22	0.10	0.14	5540
Identity Theft	0.36	0.73	0.48	2921
Intimidation	0.35	0.32	0.33	2250
Motor Vehicle Theft	0.24	0.29	0.26	4563
Robbery	0.12	0.08	0.10	1319
Shoplifting	0.32	0.62	0.42	2975
Simple Assault	0.47	0.56	0.51	3790
Theft From Building	0.09	0.09	0.09	1492
Theft From Motor Vehicle	0.30	0.24	0.26	7111
Theft of Motor Vehicle Parts or Accessories	0.17	0.17	0.17	4061
Trespass of Real Property	1.00	1.00	1.00	1711
accuracy			0.32	52565
macro avg	0.32	0.35	0.33	52565
weighted avg	0.30	0.32	0.30	52565

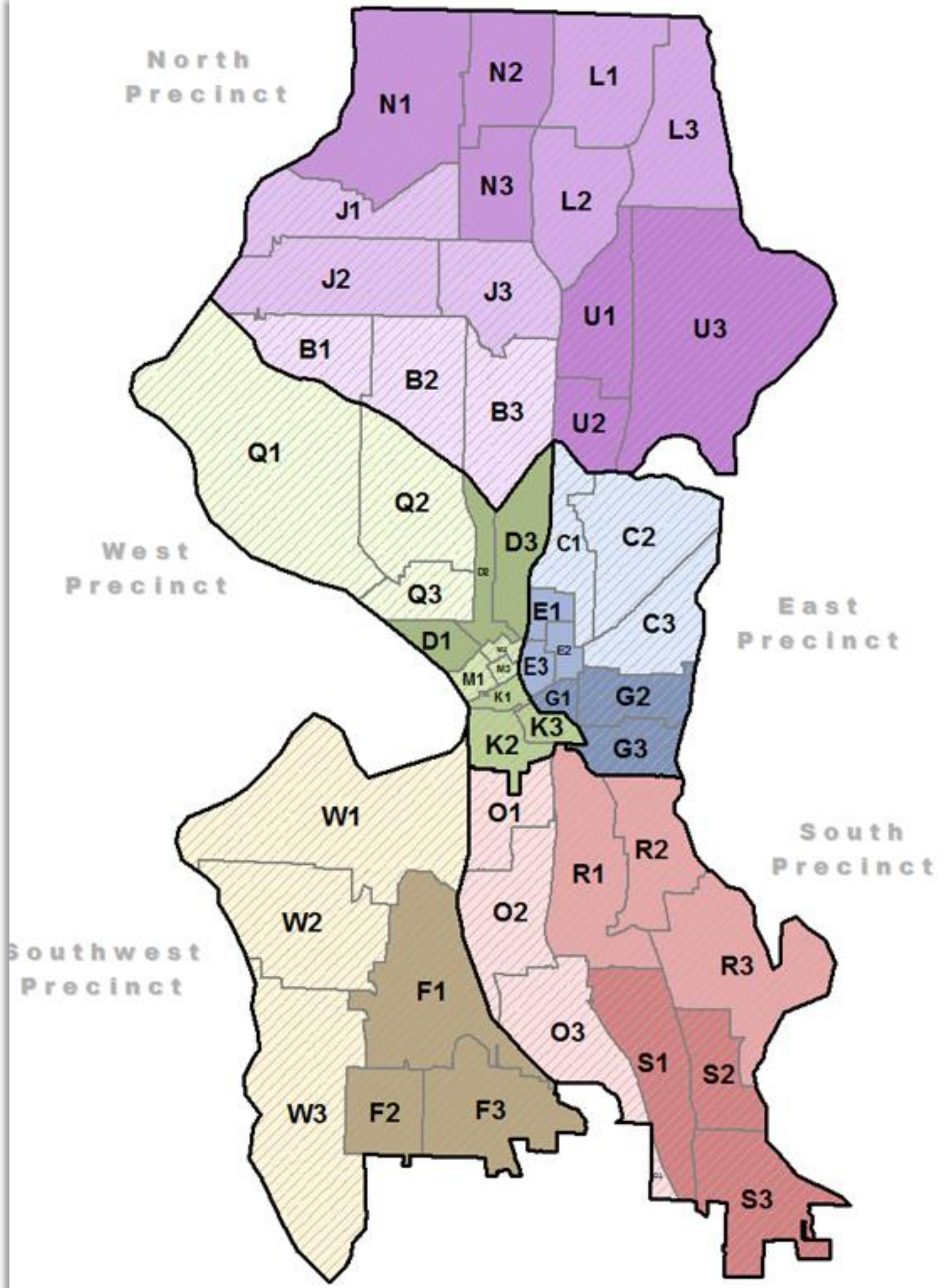
Crime Type TP/FN

- Shoplifting had 1,859 true positive instances, 1,116 false negative instances
- Simple Assault had 2,139 true positive instances, 1,651 false negative instances



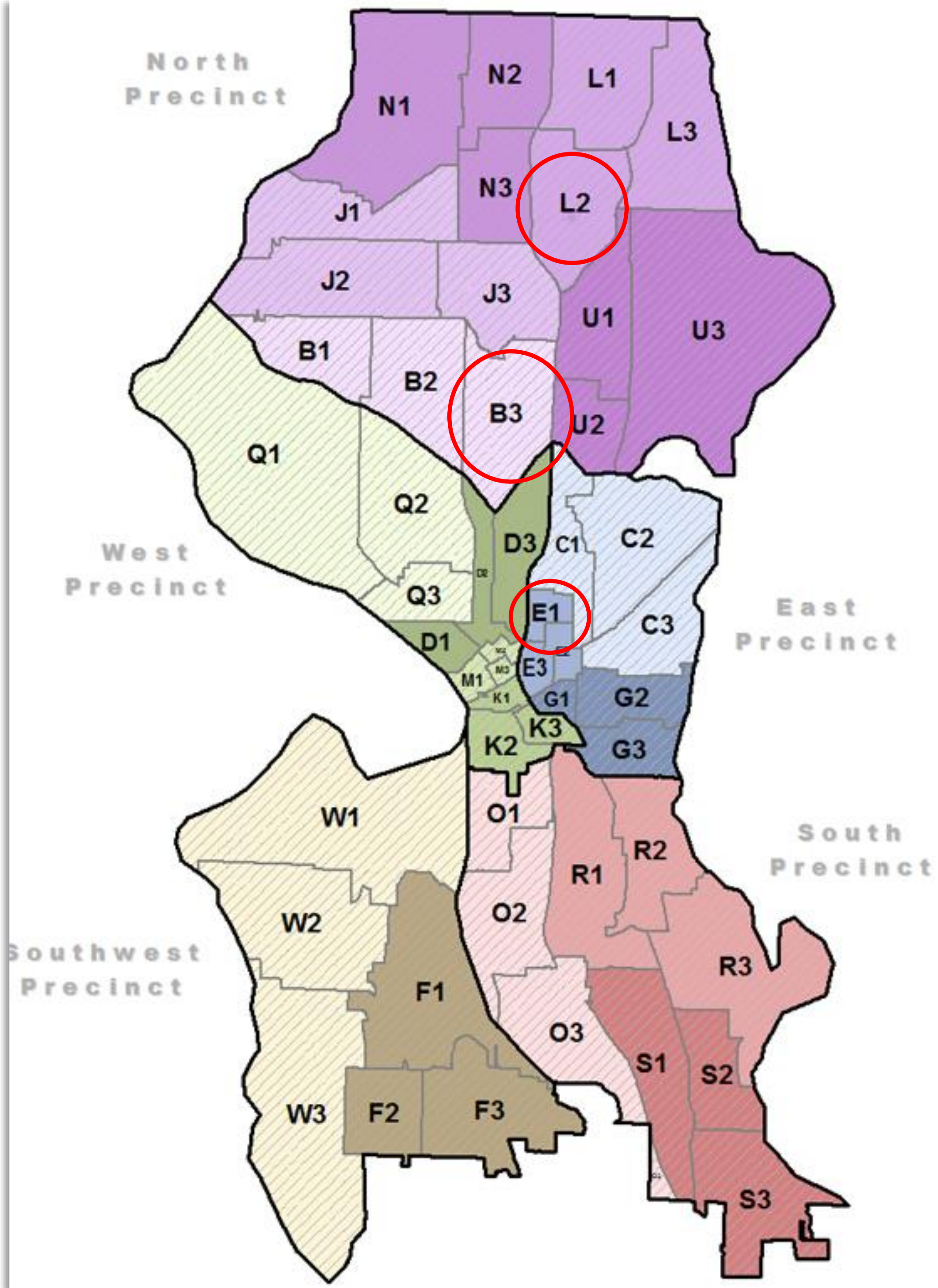
Crime Location Results

- Seattle separated into precincts, sectors, and beats
- Beats are what we use
- Recall, TP/FN



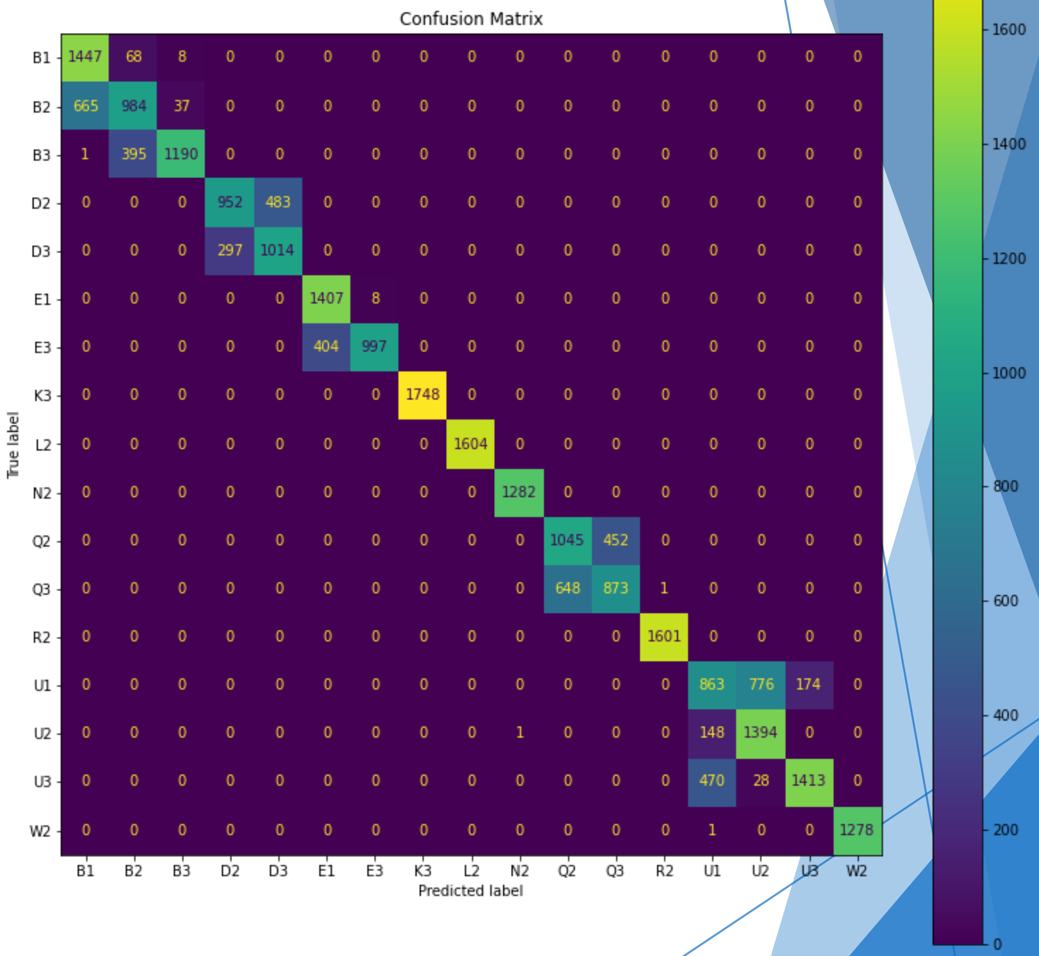
Crime Location Recall

- Best predicted crime locations were B3, E1, L2 Beats
- For recall B3 was .75, E1 was .99, and L2 was 1.0
- High recall



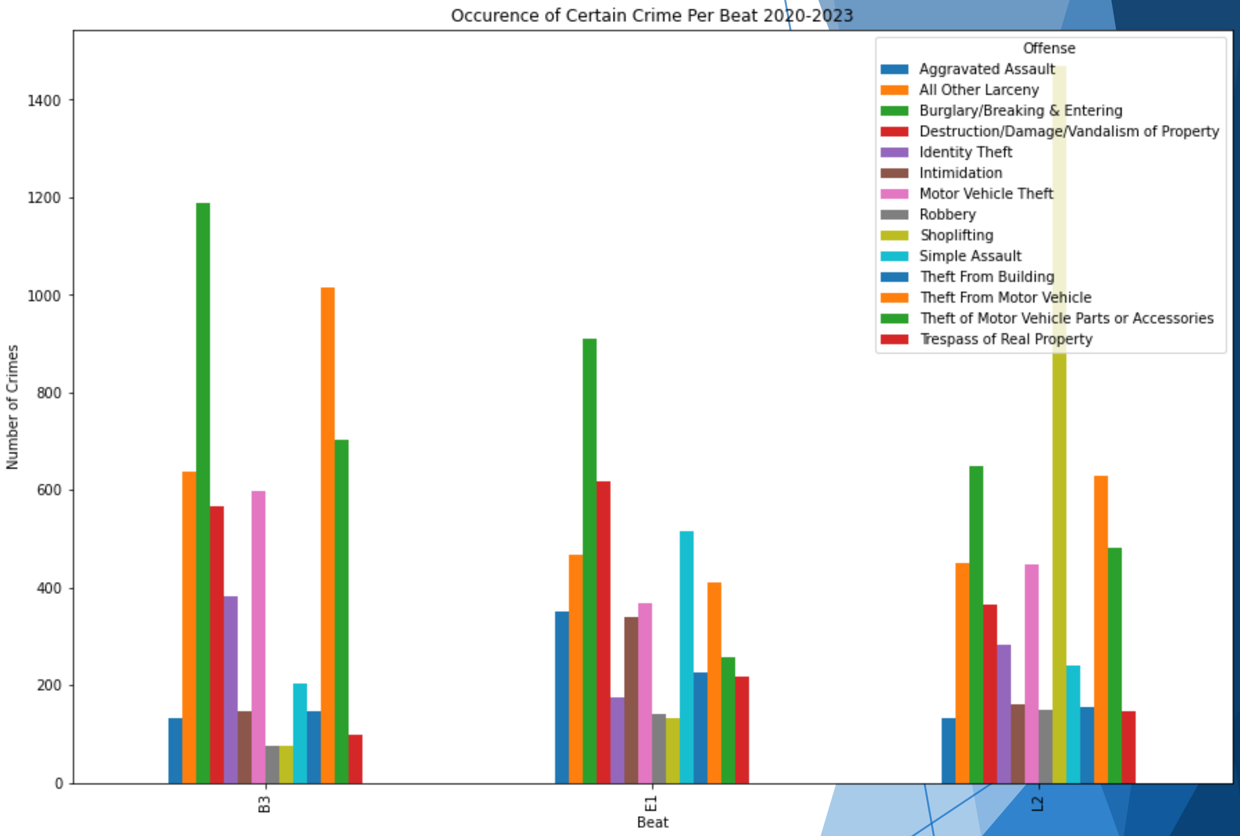
Crime Location TP/FN

- For True Positives, B3 had 1,187 instances, E1 had 1,406 instances, and L2 had 1,604 instances
- For False Negatives B3 had 399 instances, E1 had 9 instances, and L2 had 0 instances.



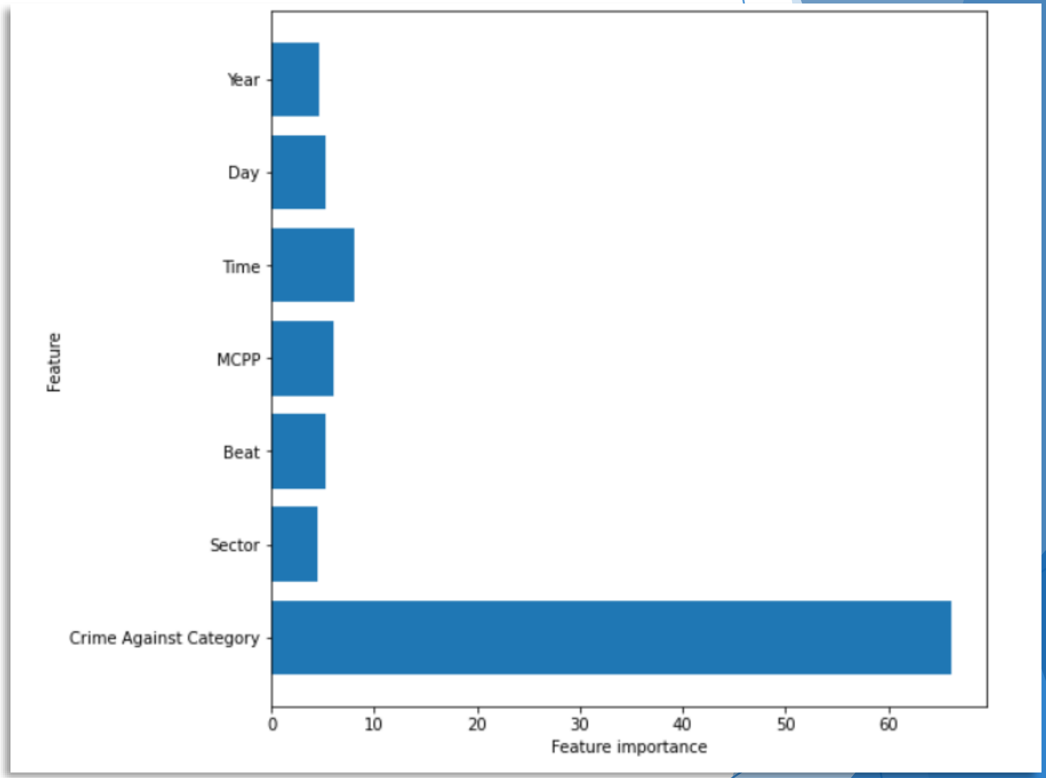
Additional Analysis

- Looked at data without the use of ML models
- Looked at which crimes occurred the most in our best predicted Beats
- We can see top crimes between the 3 Beats are Burglarly/Breaking & Entering and Theft From Motor Vehicle crimes



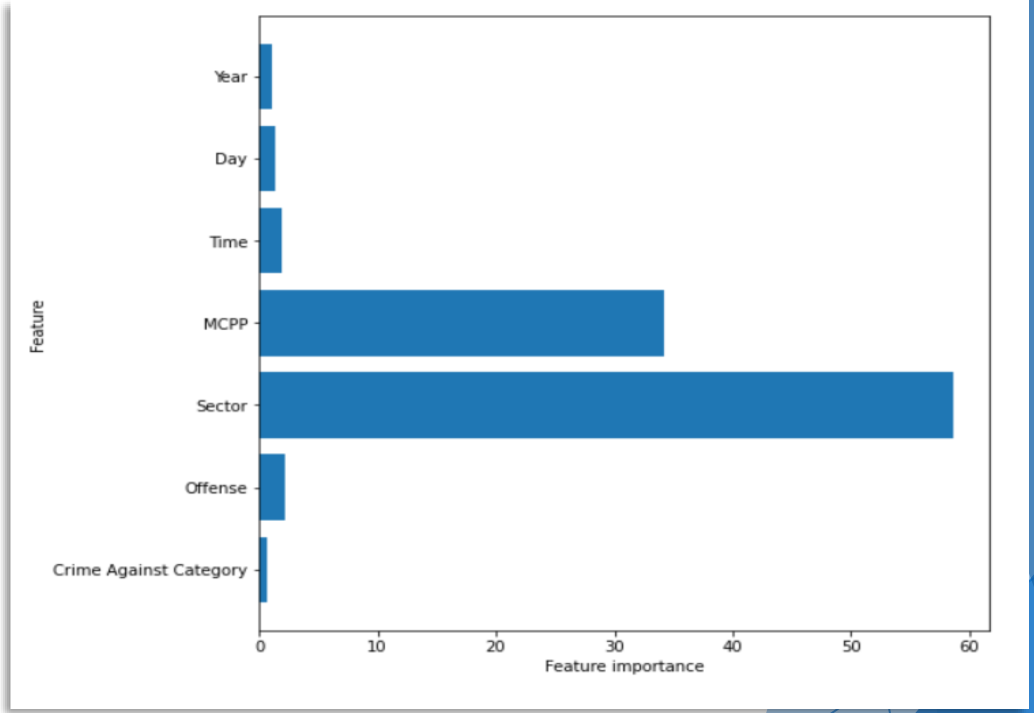
Crime Type Feature Importance

- The variable that is most useful for predicting crime type is the “Crime Against Category” variable
- Next let’s look at crime location



Crime Location Feature Importance

- The variable that is most useful for predicting crime location is the “Sector” variable
- Look at final results

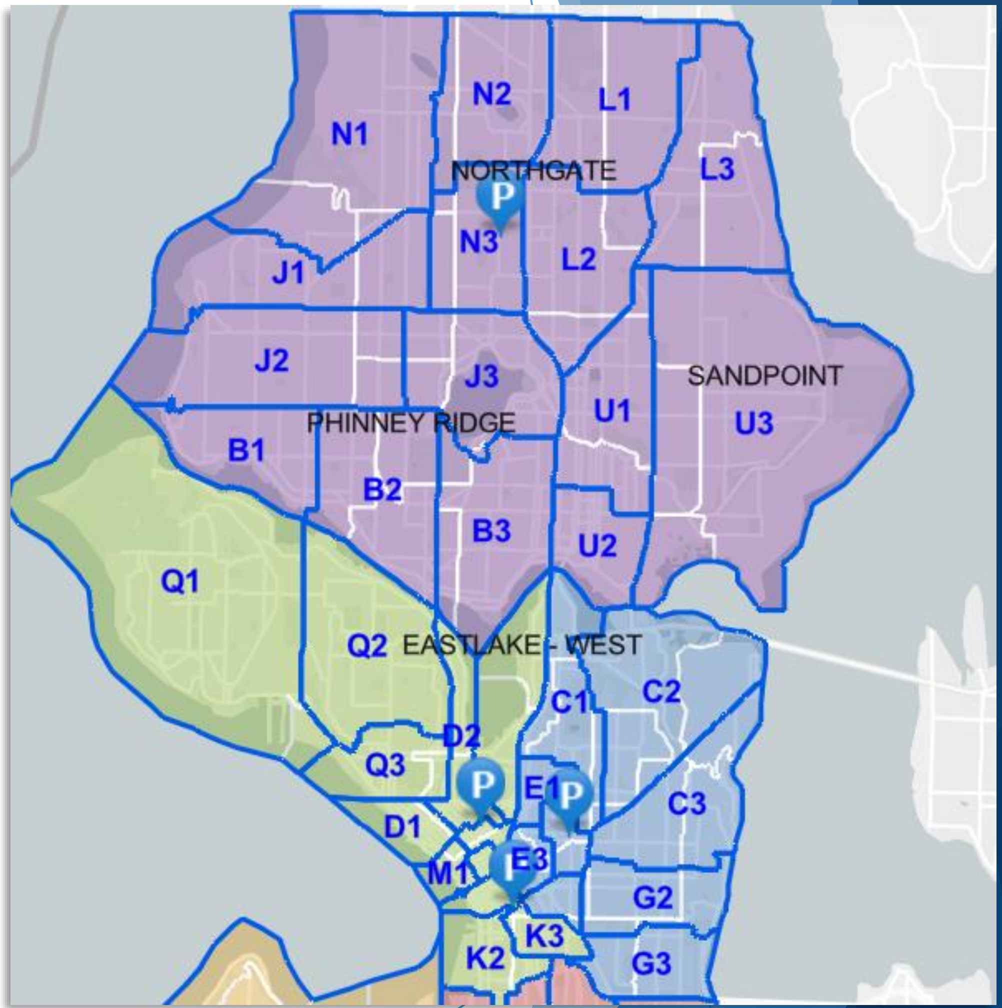


Final Results

- Our model best classifies new crimes as occurring in the B3, E1 and L2 Beats
- Top crimes are Shoplifting, Simple Assault, Burglary/Breaking & Entering and Theft From Motor Vehicle
- Variable most useful for predicting crime type is "Crime Against Category".
- Variable most useful for predicting crime location is "Sector".

Recommendations

- Initial recommendations based on what current model best predicts
- Once we improve models, able to make more specific preventative recommendations.
- Build more stations between B3 and L2
- Alternatively divert some southern police forces north
- Another option to form neighborhood watch for each Beat



Recommendations Continued

- Install more security cameras in strategic places, with 24-hour surveillance
- Easier to track all crimes, may also deter crime
- Lastly, build more monitored, affordable parking

Recommendations Continued

- Implementing model
- Implement model in Dispatch center and police laptops
- Integrate with social media

Next Steps

- Reduce overfitting in location model
- Obtain higher recall scores in crime type model
- Model takes while to train
- Run faster by training on GPU

Next Steps

- Once model successfully tuned and implemented, will have numerous positive effects
- Decrease crime, improve police-public relations, increase business/tourism
- Once proven successful, model may even be implemented in other parts WA, and then even other states

Questions and Contact Information

- Questions?
- Any additional inquiries can be directed to linkedin:
  - https://www.linkedin.com/in/alejandro-harrison/