Predicting Crime Categories For SF City and SF Districts

Vishweshkumar Patel - 012461371
Dennis Simon - 007742215
Varun Shah - 010823657

Introduction to Problem and Objective

- The project focuses to use past records of crime incidents in San Francisco to predict(classify) danger of specific type of crime occurrence at specific location of the district for certain day of week and time.
- The outcome of the project is to predict(classify) potential dangers/crimes (e.g. assault, battery, theft, drug use, etc.) for a respective district.
- The use-case provides more insights to police departments to make certain decisions to improve public safety for a respective district.
- Test three different methods for supervised classification: XGBoost, Multilayer Perceptron (Neural Network based), Ensemble
- Compare global city level model with 10 district level model

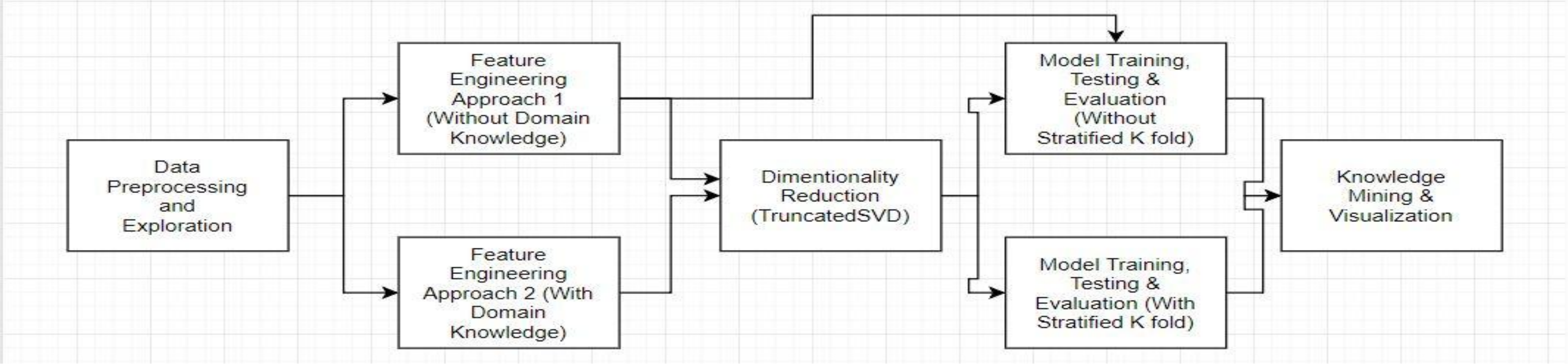
Dataset

- Dataset from ‘DataSF’ for Police Department Incidents
  - https://data.sfgov.org/Public-Safety/-Change-Notice-Police-Department-Incidents/tmnf-yvry
  - 2,206,399 total rows/incidents as of 5/1
  - 13 features

Attribute	**Type**	**Attribute Description**
IncidntNum	Number	Incident number reported for a crime
Category	Text	Category of a crime
Descript	Text	A brief description of crime incident
DayOfWeek	Text	Day, when crime happened
Date	Date	Date, when crime happened
Time	Time	Time, in between “00:00” to “23:59”
PdDistrict	Text	Police Department District
Resolution	Text	Police action(s) for respective crime
Address	Text	Apt and Street address where crime happened
X	Number	Longitude
Y	Number	Latitude
Location	Location	Location
Pdid	Number	Unique Identifier for use in update and insert operations

MLP Classifier Methodology/Architecture

- Architecture



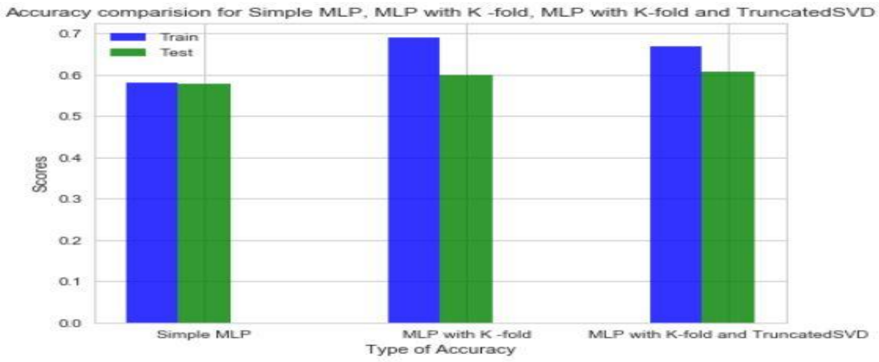
- Two approaches:
  - Approach 1 - Without considering domain knowledge ~20% accuracy
  - Approach 2 - Considering external domain knowledge to generalize categories (Index-More Serious, Non Index- Less Serious) and perform hotspot analysis ~68% accuracy with focused crime locations, more efficient to optimize resource allocation

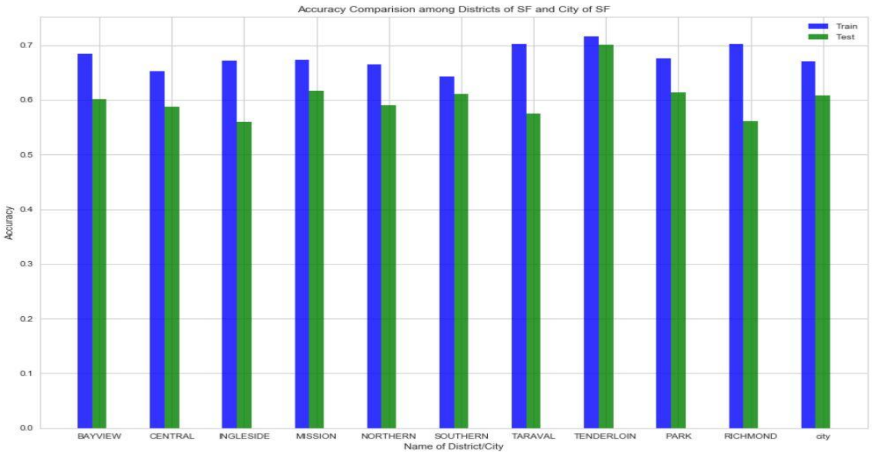
Multilayer Perceptron Classifier Results

- Improved performance using domain knowledge

- Models ‘Simple MLP’, ‘MLP with K fold’ and ‘MLP with K fold and TruncatedSVD’ exhibits nearly similar behaviour.

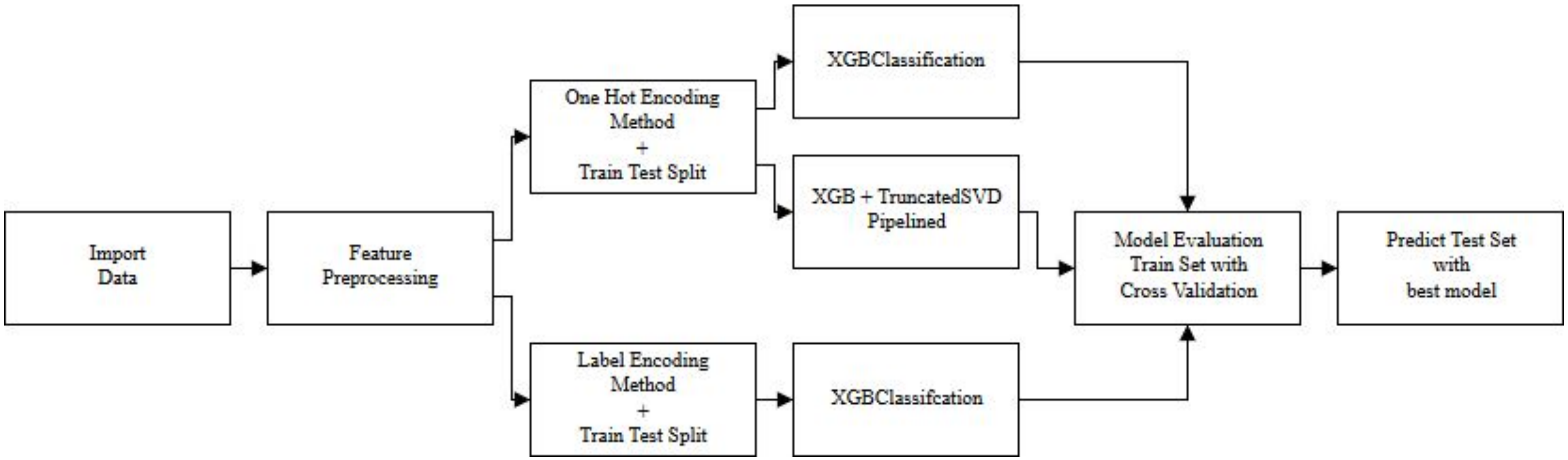
- No major difference among City level model(last bar) and district level models.





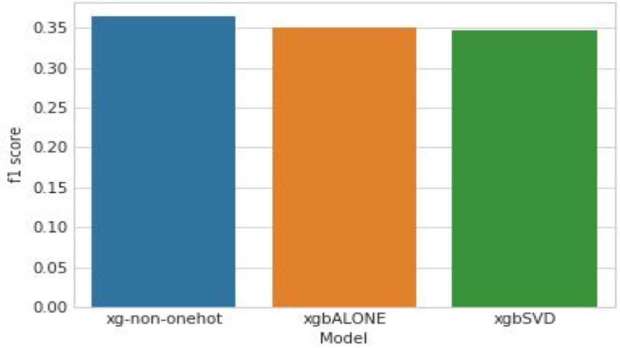
XGB Methodology/Architecture

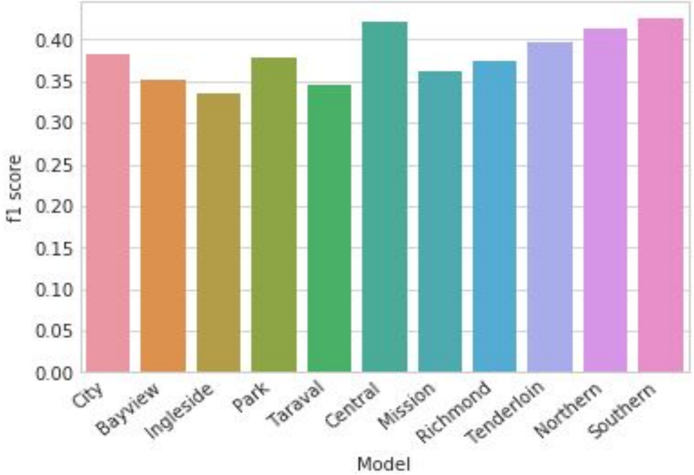
- 3 modelling methods:
  - XGB on One Hot Encoded features (sparse binary columns)
  - XGB on Label Encoded features (singular ordinal column)
  - XGB w/ Truncated SVD dimension reduction on One Hot Encoded features
- Tune hyperparameters using RandomizedSearchCV and StratifiedKFold



Extreme Gradient Boosting (XGB) Results

- 3 Methods had extremely similar f1 scores
  - XGB on one-hot-encoded marginally faster
    - Applied to all districts and whole city
- F1 scores similar throughout different districts/city
- Low scoring/results, attributed to lack of relevant/important features to properly categorize into 39 different label possibilities.

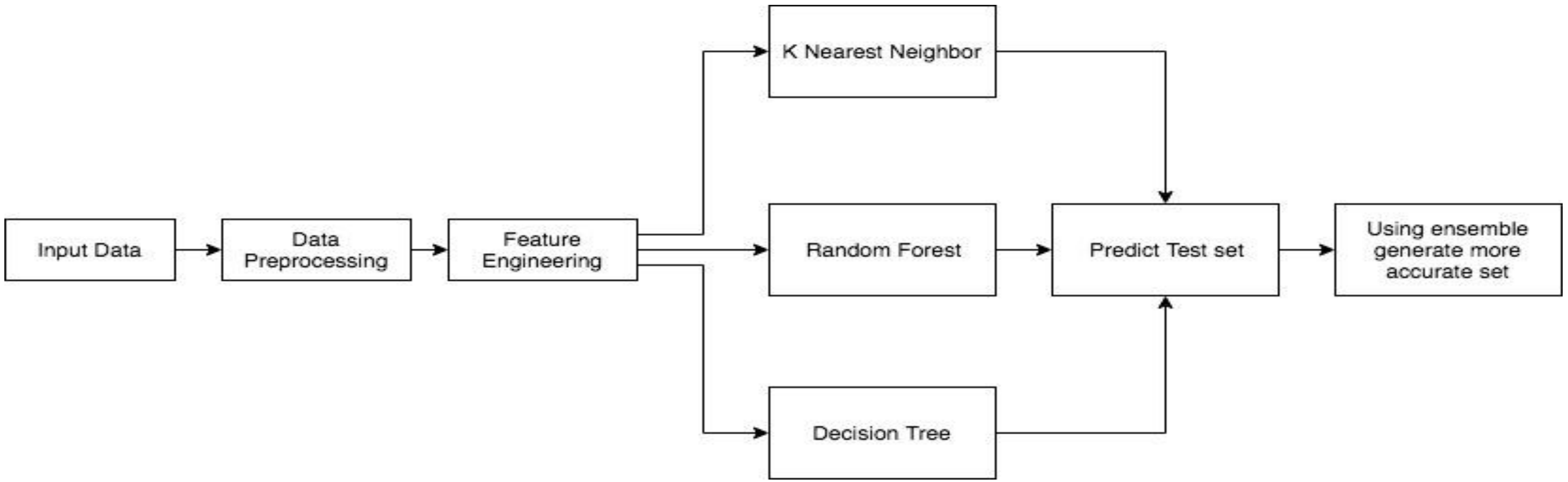




Ensemble

Ensembling of following Classification Models:

- K Nearest Neighbor
- Decision Tree
- Random Forest



Ensemble

- All 3 algorithms provides almost equivalent F1 Score
- Improvement in performance by further categorizing 39 type of crimes into Index (Severe) and Non-Index Crime (Less Severe)
- Ensembling algorithms provides nearly better accuracy