San Francisco Crime Classification

Suhas Gupta
Neha Kumar
MIDS 207 Spring 2019

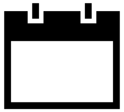
Introduction of Problem

[(Kaggle Link)]

Predict the type of crime given the time and location

- City of San Francisco
- 1934 to 1963
- Supervised learning
- **Classification** problem

Data Provided in Test and Training Dataset



Time-related Data

Timestamp of crime
Day of week of crime

Location Data

Latitude
Longitude
Address

Other Data

Police District

EDA

- ❖ Skewed categories
- ❖ Anomalies with geo-coordinates
  - ➢ 90^0 latitude (North Pole!)

Baseline Model	Log-Loss	Kaggle Rank
Predict Larceny	27.67	2151

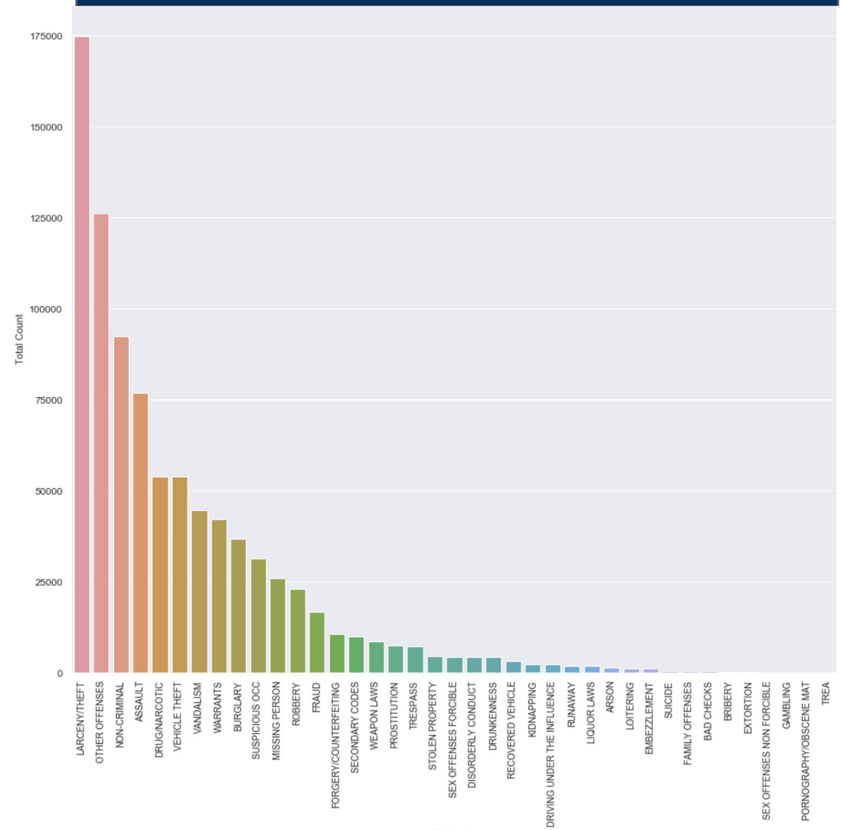
Training Data Set

Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873
2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.438738	37.771541

Test Data Set

Id	Dates	DayOfWeek	PdDistrict	Address	X	Y
0	2015-05-10 23:59:00	Sunday	BAYVIEW	2000 Block of THOMAS AV	-122.399588	37.735051
1	2015-05-10 23:51:00	Sunday	BAYVIEW	3RD ST / REVERE AV	-122.391523	37.732432
2	2015-05-10 23:50:00	Sunday	NORTHERN	2000 Block of GOUGH ST	-122.426002	37.792212
3	2015-05-10 23:45:00	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.437394	37.721412
4	2015-05-10 23:45:00	Sunday	INGLESIDE	4700 Block of MISSION ST	-122.437394	37.721412

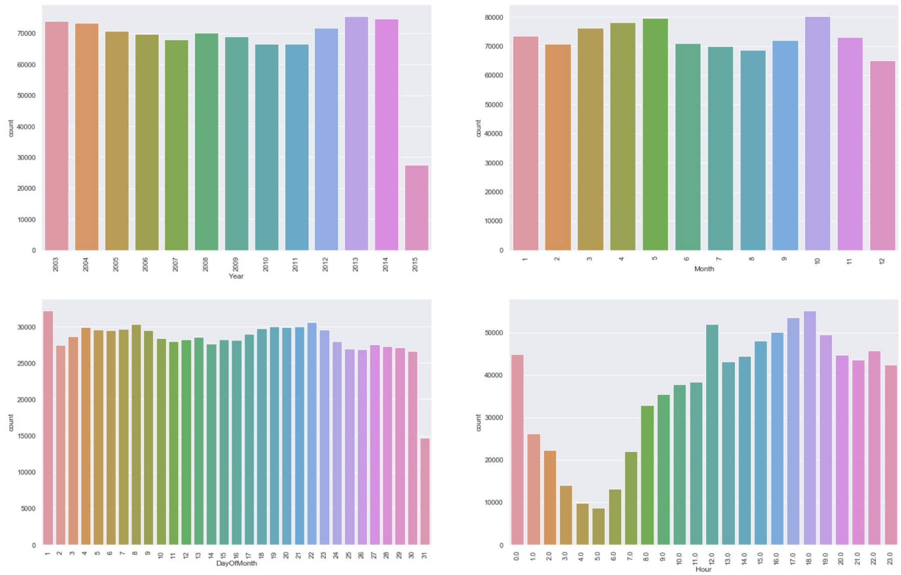
Crime Frequency by Crime Type



Feature Engineering

- Extracting date information
- Removing original timestamp
- Adding holiday flag

- Adding Zip Codes using KNN
- Normalizing Latitude/Longitude



Date extraction

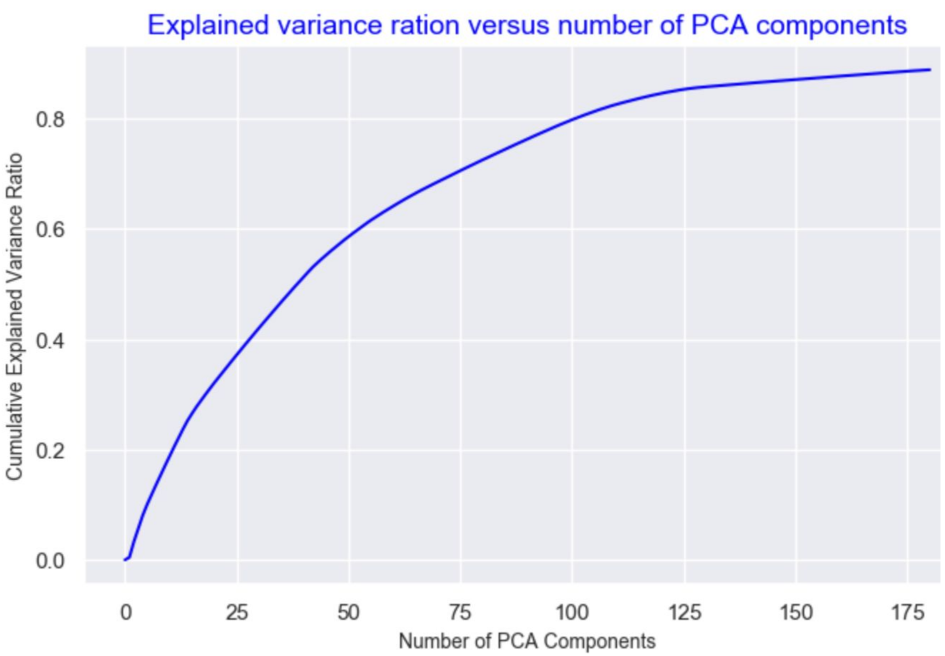
Latitude

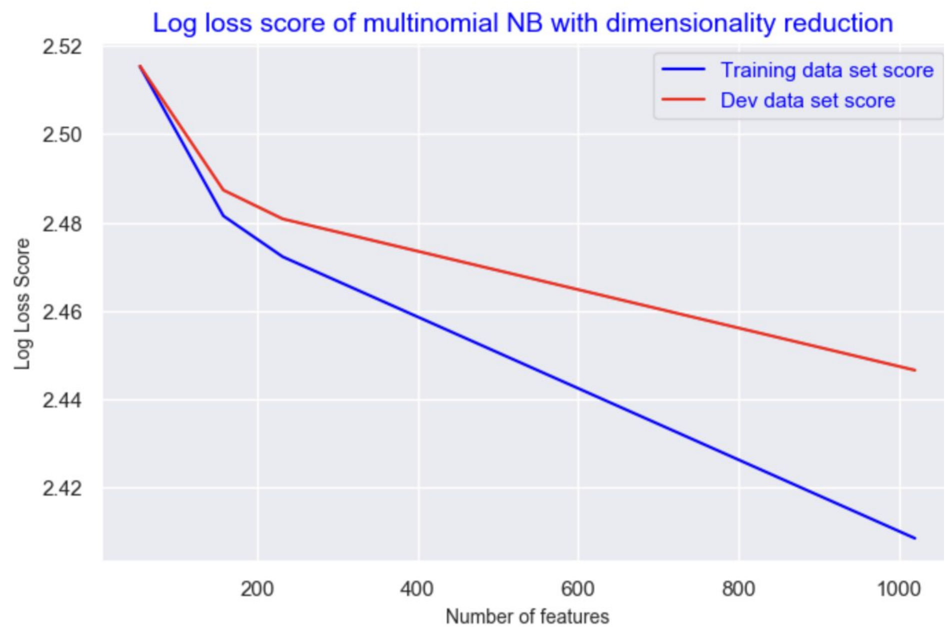
Longitude

❖ One hot encoding of all categorical variables
❖ Numerical encoding of outcome variable

Dimensionality Reduction

Feature Reduction Method	Best Training Score	Best Development Score
None (Full Feature Set)	2.208	2.563
L1 Logistic Regression	2.408	2.446
PCA	2.647	2.644



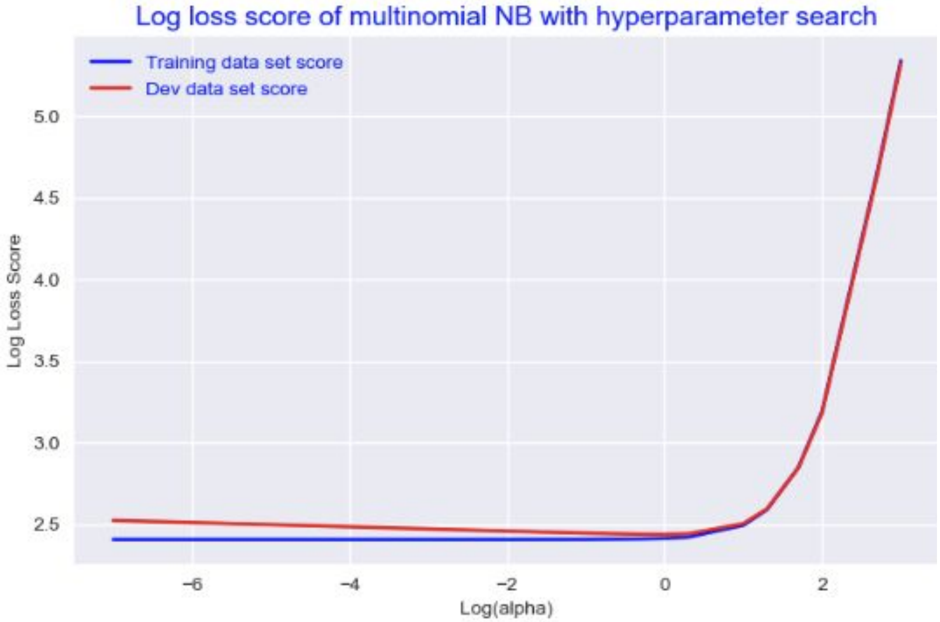


Model 1: Multinomial NB

$$\widehat{\ell}(\theta\,;x)=\frac{1}{n}\sum_{i=1}^{n}\ln f(x_i\mid\theta)$$

Model Rationale

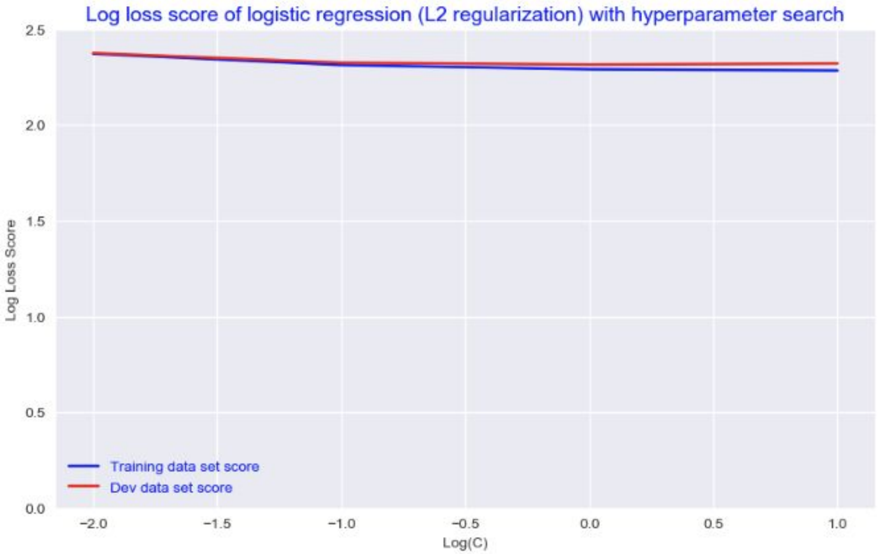
- ❖ Uses posterior conditional probabilities
- ❖ Allows for non linear decision boundaries
- ❖ Less affected by a skewed classifiers than KNN
- ❖ Quick to iterate
  - ➢ Used to assess dimensionality reduction



Model 2: L2 Logistic Regression

Model Rationale

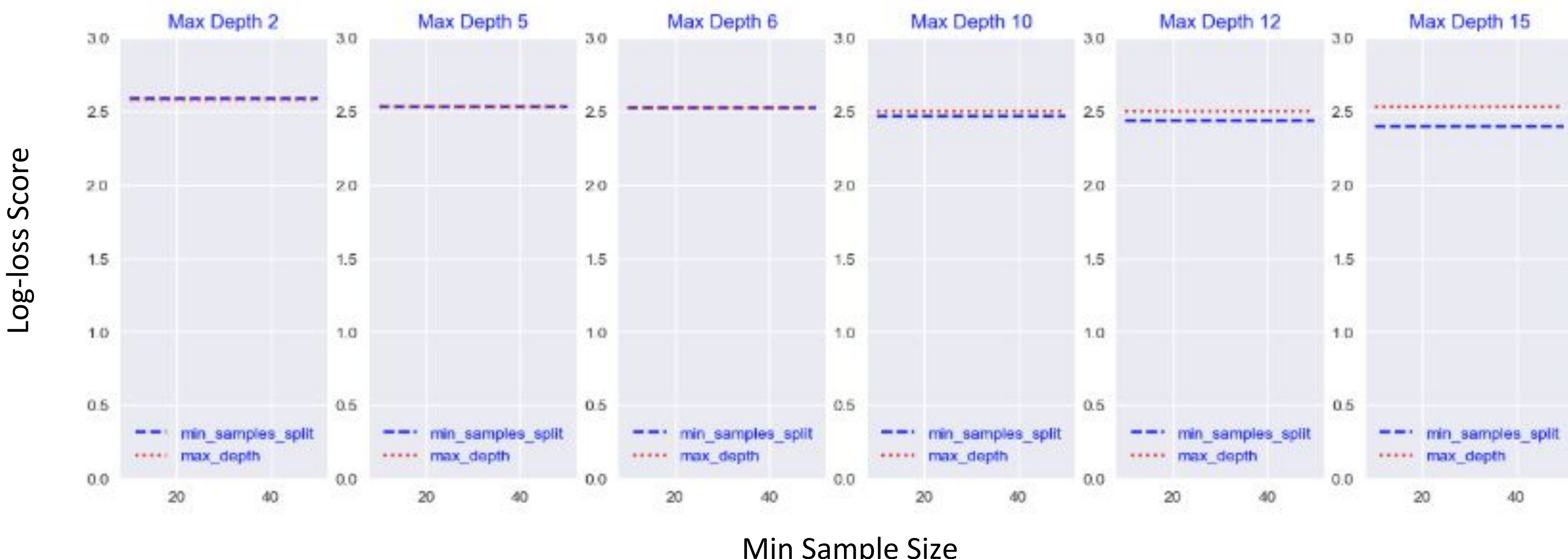
- Assumes Linear Boundaries
- Impact of dimensionality reduction
  - Dimensionality reduction allowed the model to converge
- Optimal C = 1.0



Model 3: Decision Trees

Model Rationale

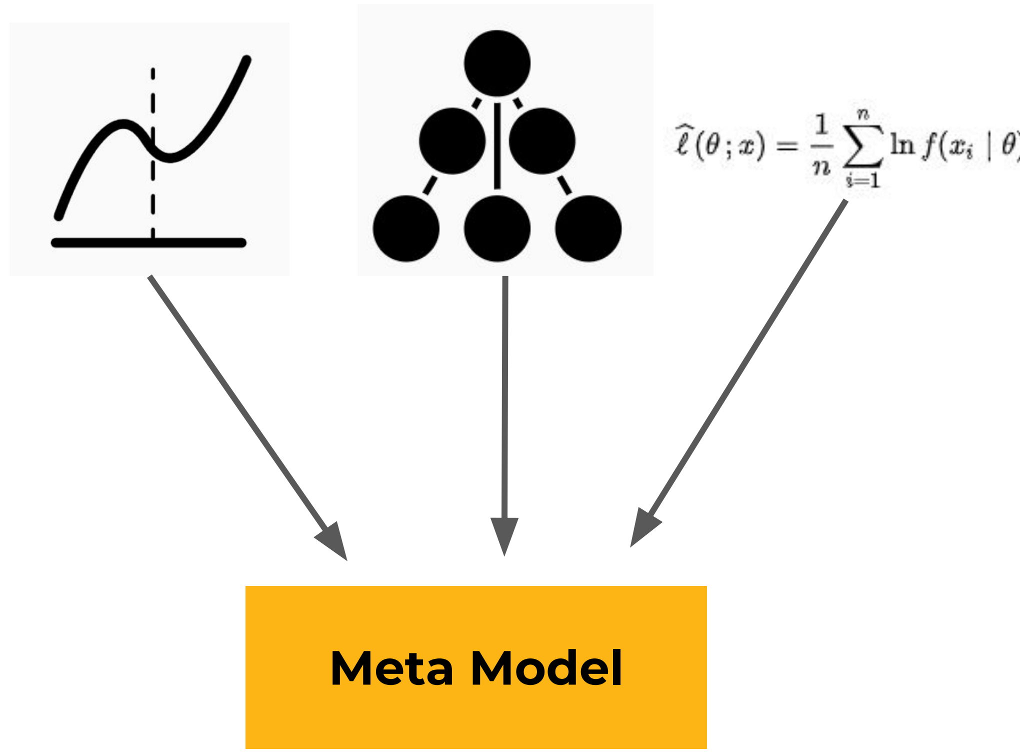
- Do not assume Linear Boundaries
- Optimal Max Depth = 6, Optimal Min Sample Size = 50 (most parsimonious)
- Advanced Tree-based methods did not improve performance compared to regular Decision Trees
  - Random Forest Log Loss: 2.52063 (dev)
  - Adaboost Dev Log-Loss: 2.68086 (dev)

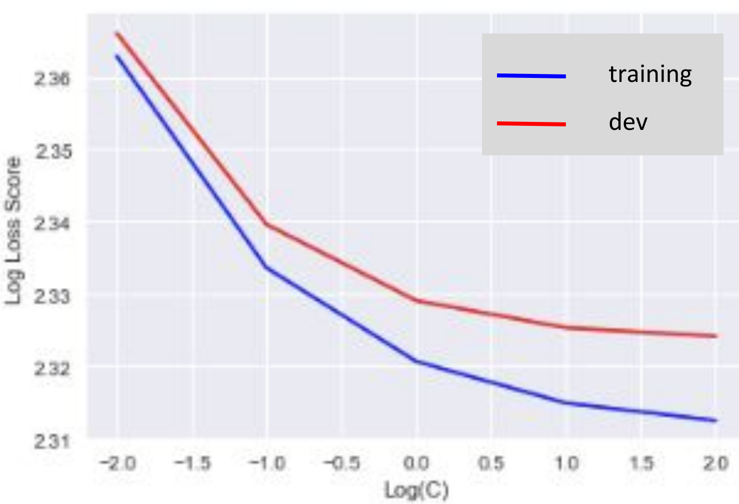


Testing a Meta-Model

Methodologies Tested

1. Averaging probability estimates
2. Using probabilities as features for a second L2 model, optimized below





Model Evaluation

Model Scores (Log Loss)

Each model performed fairly well with feature engineering
Metamodel takes the best of all 3, adding a level of nonlinearity

Model	Train	Dev	Test	Kaggle Rank
Decision Trees	2.5191	2.5231	2.52915	829
Multinomial NB	2.4146	2.4392	2.44809	655
L2 Logistic Reg	2.2914	2.3160	2.32627	404
Simple Average	2.3281	2.3429	2.34919	454
L2 Meta Model	2.3125	2.3241	2.33140	420

★ Our best model is L2 Logistic Regression only, outperforming even our two meta-models

Thank you!

Citations:

[Scikit-learn: Machine Learning in Python](), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.