Predicting Building Permit Issue Time

David Tse

Springboard Capstone Project 1

September 3, 2018

Project Motivation

- Building permit issue times
  - Potentially increases construction project schedules
  - Financial risk if projects not completed on time
  - Few general guidelines given from city building departments
  - Typically require building expeditors to hasten permit issue times

Potential Client Use Cases

- Real Estate Developers: **optimize their project portfolio** to consider reduced construction project timelines

- Homeowners: obtain more accurate time estimates for permitting and **reduce reliance on anecdotal advice** from building contractors/building department.

- NYC Building Permit Officials: **triage requests** to potentially reduce response turnaround time.

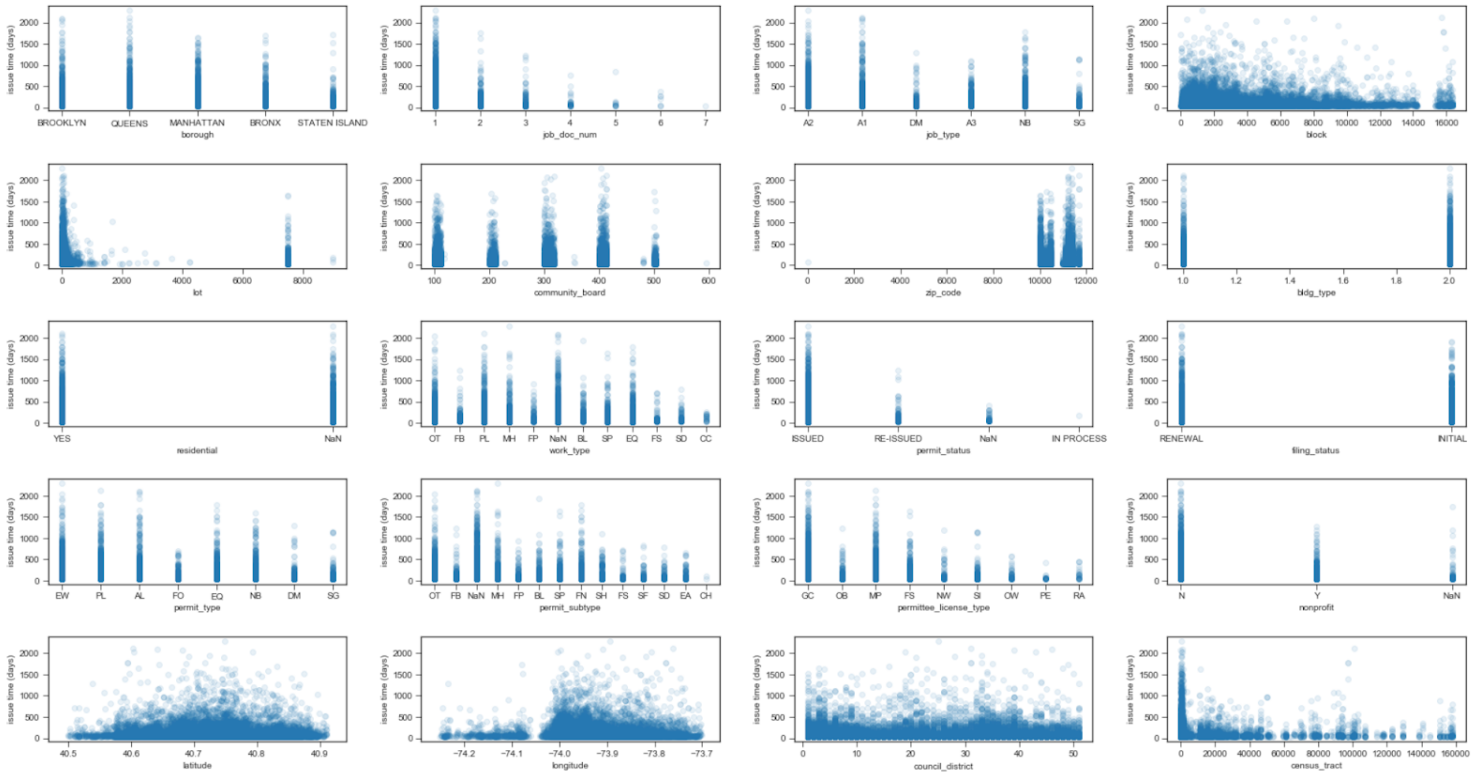
Data Wrangling and Cleaning

- Sourced from NYC OpenData
  - CSV file: 60 columns and over 3.37M records
  - Filter date range from May 9th 2012 to May 9th, 2018
- Cleaning:
  - Outliers: Removed negative issue times
  - Missing Data: Marked as NaN, or removed if # of missing records is <5% of total #
  - Out of Scope: features with extremely high cardinality (e.g. address) or optional fields with a majority of missing values
- Feature Engineering:
  - Dependent variable:
    - Continuous: Issue time (days) = issuance date - filing date
    - Binary: 1 to 3 months or over 3 months
  - Extracted day and month information from temporal data

Summary Statistics

	job_doc_num	latitude	longitude
count	17589.000000	17566.000000	17566.000000
mean	1.072091	40.726829	-73.937831
std	0.341786	0.072279	0.076697
min	1.000000	40.499227	-74.252245
25%	1.000000	40.682072	-73.985089
50%	1.000000	40.728893	-73.952962
75%	1.000000	40.768984	-73.901479
max	7.000000	40.911082	-73.701445

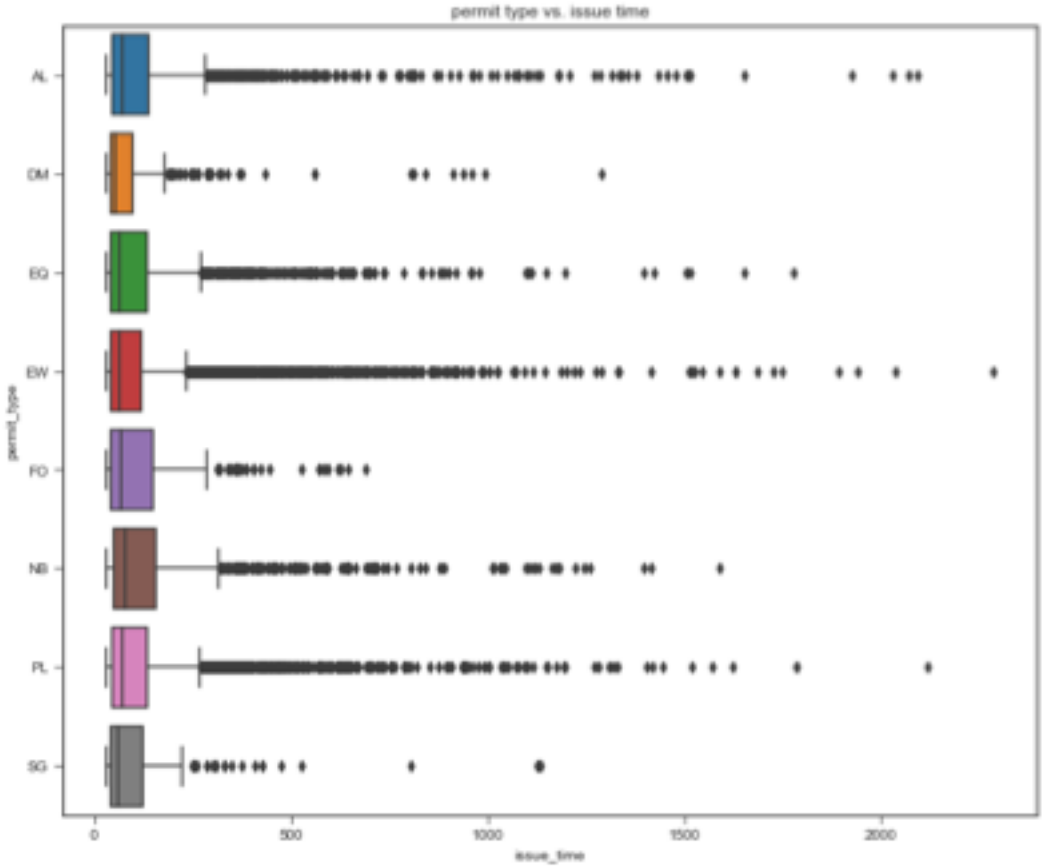
Independent Variables vs. Issue Time

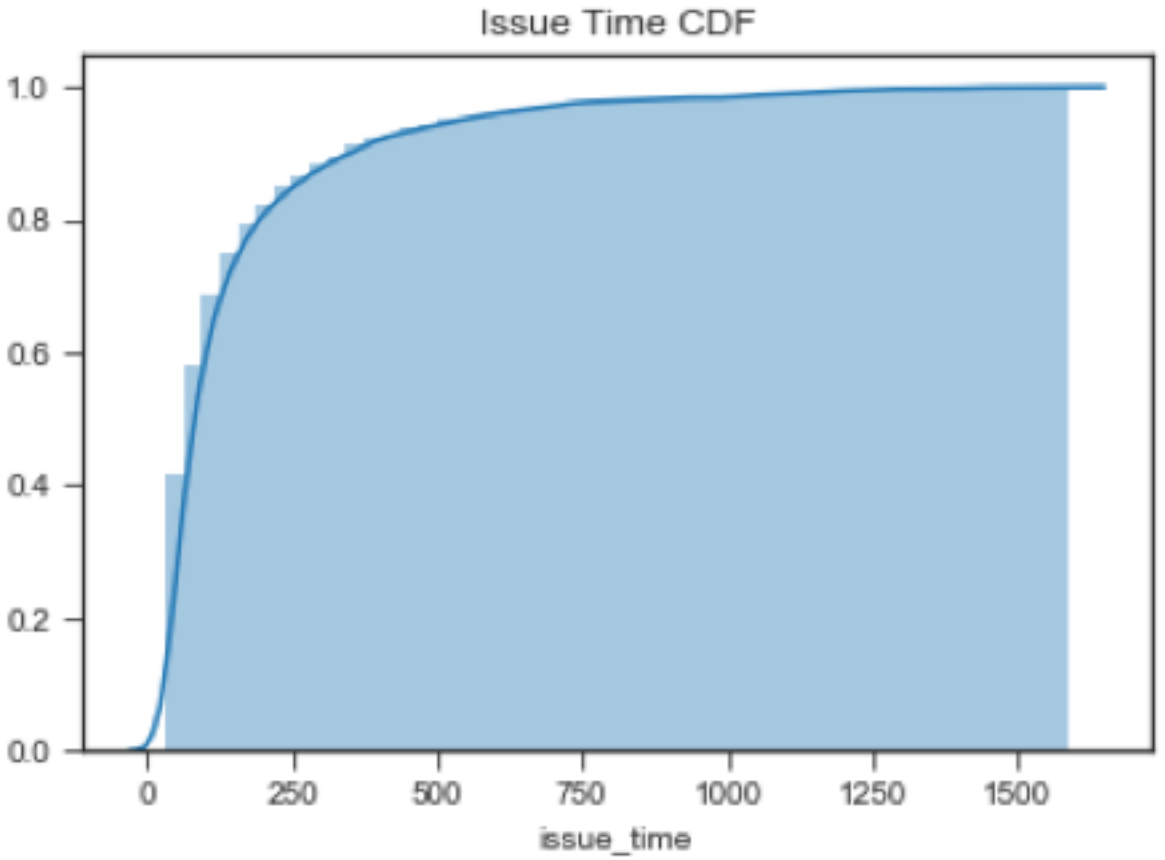


Observations:

- Non-profits have lower issue times.
- Queens had the largest spread, while Manhattan had the least.
- As the number of documents in the application increases the issue time seems to decrease (surprisingly).
- Foundation related permits are one of the fastest permit types.

Heavy Right Skewness of Data





Building Permit Issue Time Heat Map

Neighborhoods with high activity:

- **Brooklyn**: Park Slope, Bushwick, Brighton Beach, Williamsburg
- **Manhattan**: Majority of neighborhoods
- **Queens**: Flushing, Jackson Heights, Elmhurst, Astoria, College Point
- **Bronx**: N/A
- **Staten Island**: N/A

Framing the Machine Learning Problem

Classification or Regression?

Use Binary Classification as a starting point:

- fewer assumptions to be satisfy
- Still provides value to the end user
- Easier to address class imbalance issues given heavy right skew in data

Which metrics to use?

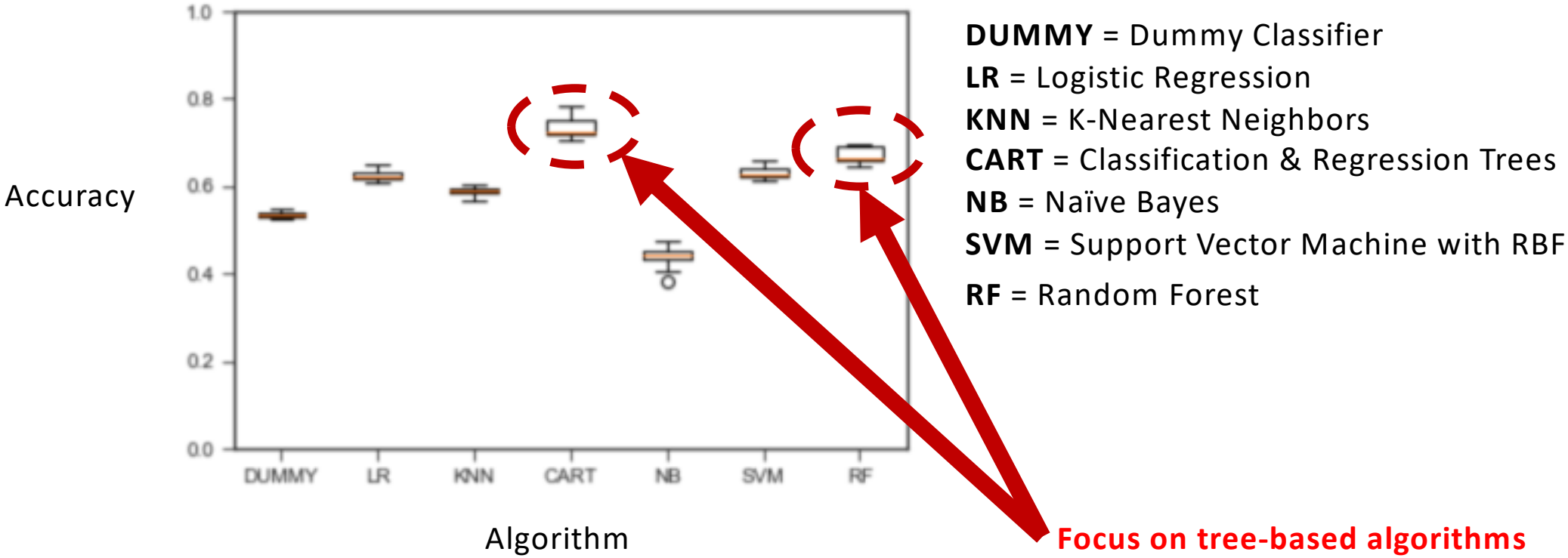
- AUC of ROC Curve
- F1-score
- Result Interpretability

Feature Importance

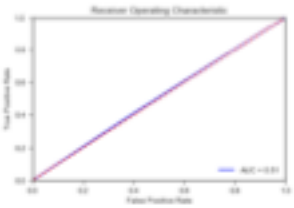
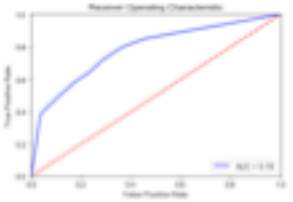
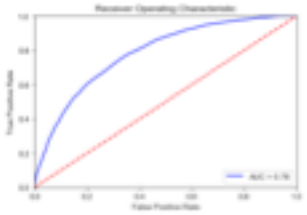
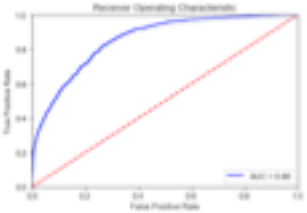
Top Feature Types:

- Temporal
- Location
- Work Type

Initial Machine Learning Algorithm Comparison



Optimized Machine Learning Models

	Optimized Hyperparameters	F-1 Score	ROC Curve AUC	
Dummy	N/A	0.53	0.51	
Decision Tree	'max_depth': 50, 'min_samples_leaf': 4	0.73	0.78	
Random Forest	'max_depth': 150, 'min_samples_leaf': 1 'n_estimators': 100	0.68	0.78	
Gradient Boosted Tree	'max_depth': 150, 'n_estimators': 200	0.76	0.86	

Provides interpretability with decision stumps, fairly high AUC and robustness against outliers

Oversampling leads to improved balance between precision and recall

Recommendations

Real Estate Developers/ Homeowners:

- 1) Location, time and work/permit type are some of the most important features in building issue times.
- 2) Use decision trees (real estate developer use case) or XGBoost (homeowner use case) to predict building permit issue time durations (medium or long duration) for non-trivial work items, e.g. new buildings and major alterations that will change the use, egress, or occupancy of the building.
- 3) For minor work, such as electrical work and demolition, involving single building departments expect issue times of less than a month.

NYC Building Department:

- Provide a web service that allows applicants to take an online survey.
- Survey generates building permit issue time range based on data from the past 5 years based on the XGBoost algorithm.

Limitations and Future Work

- Limitations:
  - The building permit data did not include the cost of the proposed work.
  - Building permit expeditors typically reduces turnaround time for permit applications, which may introduce bias into the dataset.
- Future Work:
  - Investigate framing the problem into a multi-class problem.
  - Try incorporating other related datasets that may boost the signal, such as population and household income datasets to better understand why location is a strong predictor.

Acknowledgements

- Springboard staff and community
- My Springboard mentor
- NYC OpenData