Exploratory data analysis and modeling of public building construction in New York City

Culture + Data — December 14, 2022

Sara Venkatraman
Cornell University, Statistics and Data Science
skv24@cornell.edu

Data overview

The DDC maintains a database of ~170 attributes of ~8000 public building and infrastructure projects in NYC from the last 35 years

- **Project phases:**
  Initiation → Design → Construction procurement → Construction → Closeout

Project ID	Borough	Phase	Phase Start	Phase End	Project Type	…	Budget	Sponsor	Private Funding	Demolition
227Duffield	BK	Design	7/16/2021	10/31/2021	Structural Work Steel	…	$ X_1	DCAS	0	NA
227Duffield	BK	Construction	12/1/2021	2/28/2023	Structural Work Steel	…	$ X_2	DCAS	0	NA
PV788-REN	MN	Construction	2/1/2018	5/30/2018	Misc.	…	$ X_3	Cultural Affairs	1	0
…	…	…	…	…	…	…	…	…	…	…
ACEDOS501	Citywide	Construction	8/11/2015	4/8/2016	Misc.		$ X_4	DCAS	NA	NA
AGX001LC	BX	Initiation	4/21/1996	8/21/1996	NA		$ X_5	Aging	0	1

- **Our focus:** Analyzing **construction duration** for public buildings
- **Challenges:** Identifying relevant data (projects + variables), missing values

Project types

Projects are primarily divided into:

- Interior renovations (38%)
- Exterior renovations (20%)
- Miscellaneous (20%)
- Major rehabilitation (16%)
- New construction (6%)

Within these, we can have: →

Project type	Proportion of projects
Exterior roof replacement	10.8%
HVAC upgrade/replacement	10.6%
Fire alarm upgrade	8.7%
Fuel tanks	7.8%
Electrical upgrade	7.3%
Exterior façade renovation	7.1%
Boiler upgrade	2.1%
Window replacement	1.8%
Interior plumbing upgrade	0.9%
Fire sprinkler upgrade	0.5%
…	…

Cluster analysis

- **Objective:** try to automatically identify variables along which construction projects group together

- **Methodology:**
  - Apply hierarchical clustering algorithm to filtered dataset of public buildings projects in “construction” phase
  - Examine the projects in each cluster: in what ways are they similar?

- **Findings:** Groups formed around **project type, program unit, boroughs**
  - E.g. courthouse projects in Manhattan and Brooklyn with larger budgets
  - E.g. health department projects across all boroughs
  - E.g. interior renovations projects sponsored by DCAS
  - E.g. interior and exterior renovations on libraries sponsored by NYPL
  - However: within-cluster **variance** in construction duration **remains high**

What influences construction duration?

- **Correlations** between construction duration and other attributes (budget for each phase, presence of hazardous materials, demolition required, density of surrounding area, unique safety requirements) were fairly weak

- **Multivariable linear regression models** of construction duration as a function of other attributes indicate:
  - **Construction budget** has a statistically significant impact on duration
    (is budget an indicator of project *complexity*?)
  - **Project type** also has a significant effect: new construction requires most time, exterior renovation requires least
  - Impacts of **program unit** and **sponsoring agency** are less clear
    (some significant differences between program units, e.g. courthouses require more time than libraries)
  - **Borough** also has an unclear impact on construction duration, but **citywide projects** take significantly longer

Towards modeling construction durations

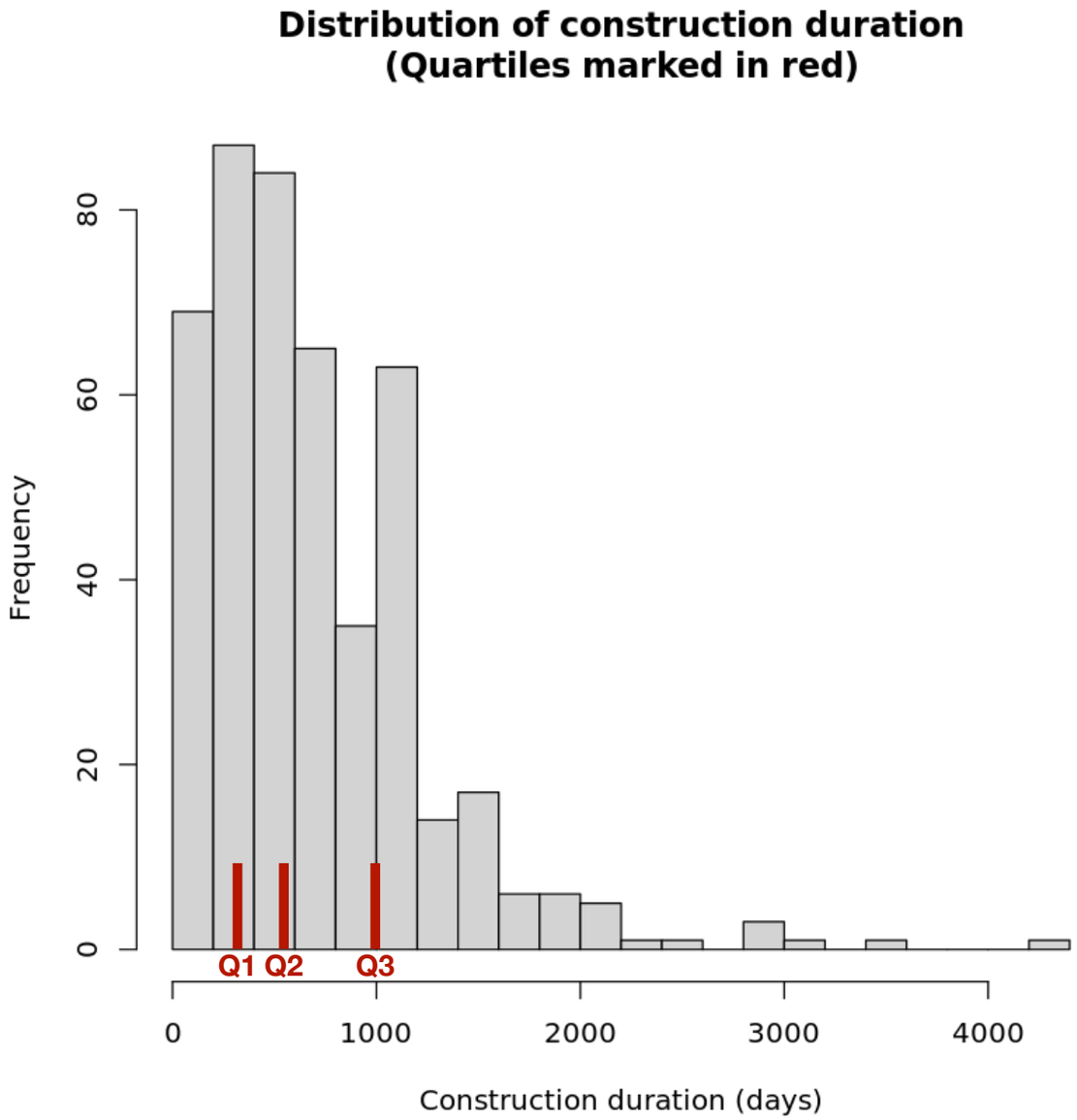
Conclusions from previous analyses:

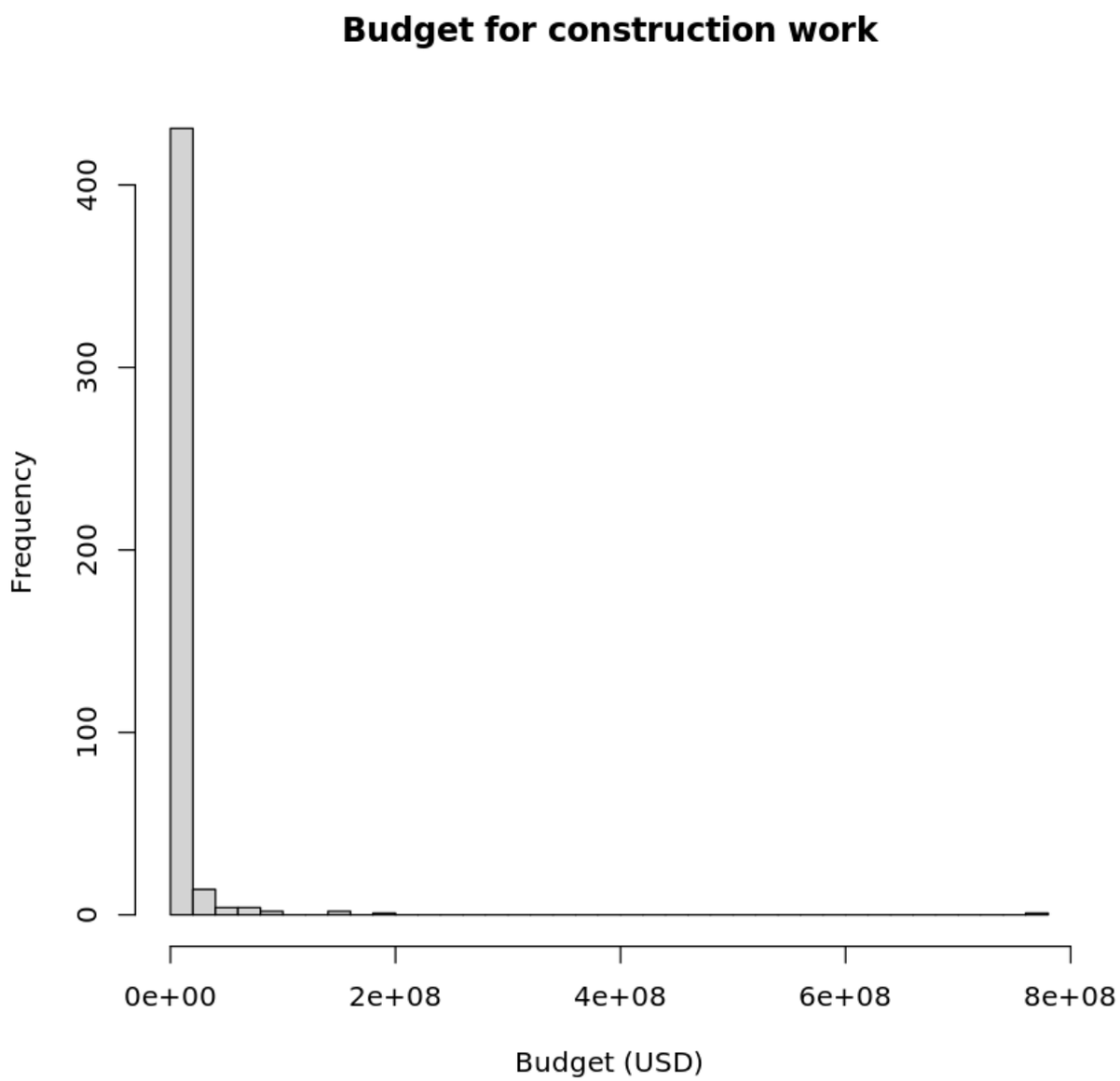
- Several variables seem predictive of construction duration
    - Construction budget (and potentially budgets of other phases)
    - Program unit and sponsoring NYC agency
    - Type of project (at level of interior, exterior, or new)
    - Location (borough-specific, or citywide)

- However: these features may affect duration in complex, nonlinear ways

  Predictive accuracy from a small number of attributes may be limited; there is a tradeoff between accuracy and interpretability of models.

Towards modeling construction durations

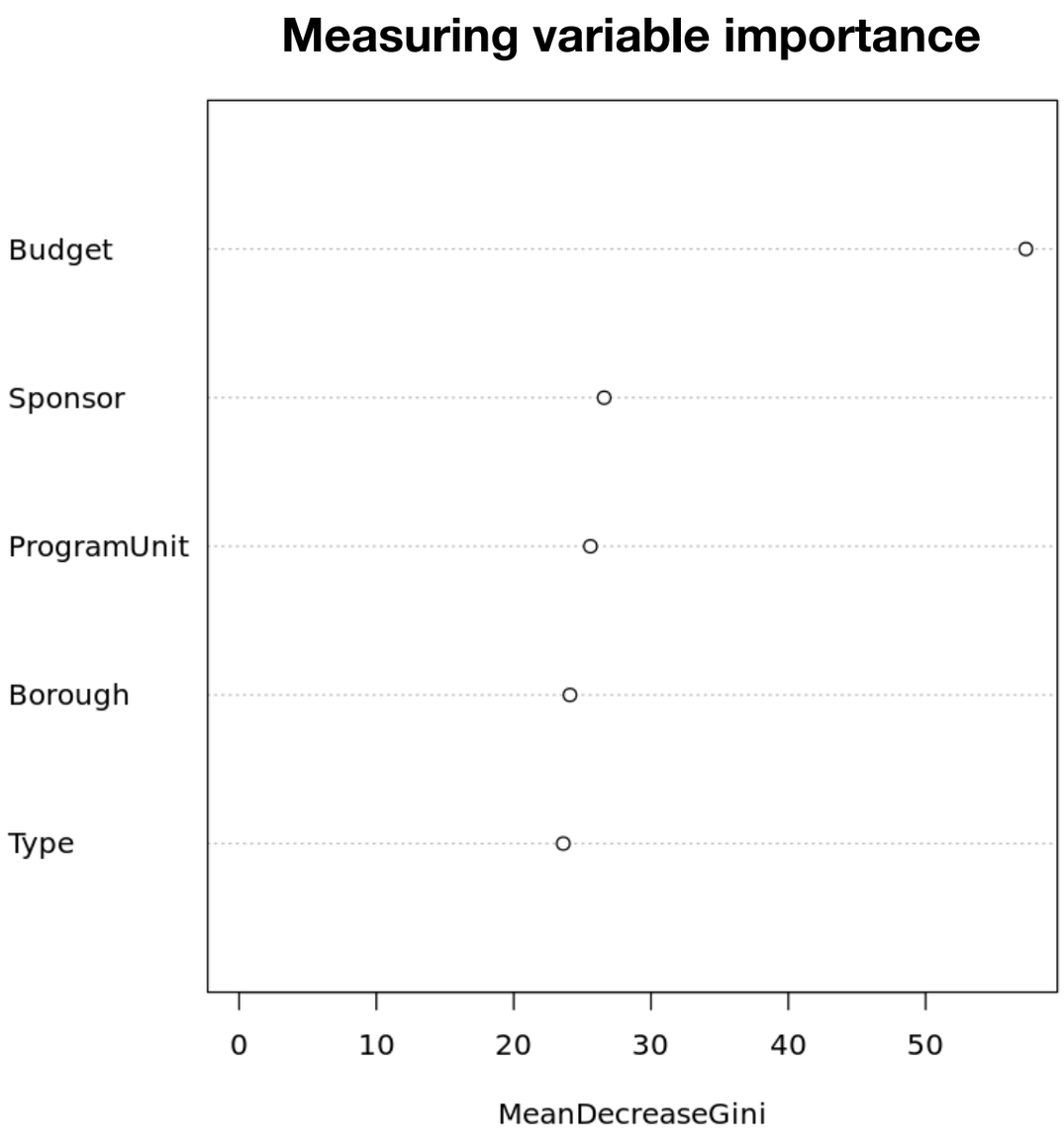




Modeling process

- **Split dataset** of public buildings into training (75%) and test (25%) sets
- **Construct random forest model of construction duration:**
  - Predictors: budget, project type, program unit, sponsor, borough
  - Two possible models:
    - **Regression model:** predicts construction duration in days
    - **Classification model:** predicts the duration interval to which a project belongs (intervals pre-defined by quantiles)
  - Classification error was highest for mid-length (~ 1.5 - 3 years) projects, so separate models were constructed for short-term and long-term projects to improve accuracy

Modeling process



Potential extensions to modeling process:

- Transformations of highly-skewed continuous predictors, and revised grouping of categorical predictors
- Expanded set of predictors with external data sources:
  - Population density
  - Traffic density
  - ZIP codes
- Quantitative measure of project complexity
- Missing data analysis/imputation