📊 Project Summary for Final ML/AI Capstone Project with Berkely Hass

Shan Gao

1 Problem Statement

What problem are you trying to solve?
What larger issues do the problem address?

Business goal: Use ML classification techniques to predict which 311 cases are most likely to be overdue.

Business outcome: Transform 311 operations from broad, generalized efforts to highly targeted and efficient operations.

See the readme for more details

2 Outcomes/Predictions

What prediction(s) are you trying to make?
Identify applicable predictor (X) and/or target (y) variables.

As Step #1 describes, predict if a new 311 call is prone to be overdue.

Label/Target Variable: 'is_overdue'

Predictors: from +100 raw features to 30 selected features for final modeling

3 Data Acquisition

Where are you sourcing your data from?
What is the dimension? Any missing values?

Variable	Value
Data Source	3 data sources
Rows	163545
Columns	30 (Final)
Missing Values Present	0%
Imbalanced?	Yes

Major Actions and Challenges:
Clean 3 data sources:
Study and solve the unit of analysis issue across the data sources

4 Data Preparation

What do you need to do your data in order to run your model and achieve your outcomes?

EDA: Descriptive analysis and plotting of variables, i.e., bar charts, histograms, boxplots, correlation Parallel Categories Plot. Cumulation

Outcome: dropping any insignificant variables.

Preprocessing Techniques:
K-mean cluster, NLP text analysis, Feature Engineering: Encoding, Binning, scaling, etc.

5 Modeling

What models are appropriate to use given your outcomes?

Modeling: Start with Random Forest to help identify important features and feature importance analysis
Modeling for All: (RF), (LR), (KNN), (DT), and (SVM)
Separate Modeling for 4 neighborhood clusters: 5 classifiers listed above.
Model Tuning and Selection: Grid Search and 5-fold Cross-Validation

Outcome(Best Models): K-nearest neighbor (all using accuracy) and Logistic(Neighborhood cluster using AUC)

6 Model Evaluation

How did you evaluate your model's performance?
Results?

A table listing the following criteria:
Classifier name, optimal threshold, accuracy, precision, recall, F1-Scire, Auc-Roc, computing time, and Hypermeter for the best model.

A ROC curve plot for all classifiers explored