Final Report
Risk and Cost Management Analysis

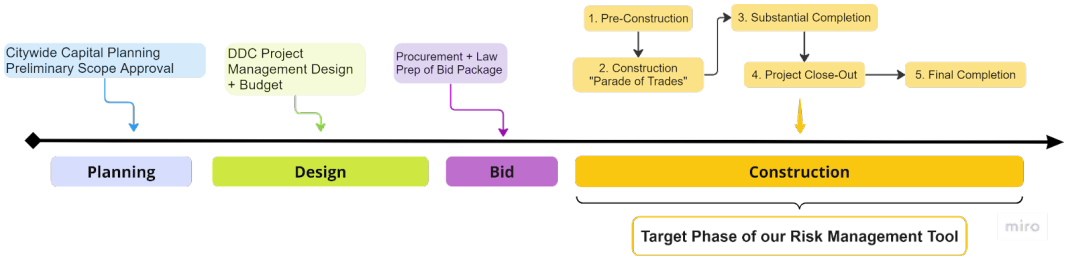
Yaoyuan Zhang (yz4387), Tongni Chen (tc3243), Zimu An (za2323)
Tao Yu (ty2479), Yijia Wang (yw3936)

July 22, 2024

1 Project Introduction

1.1 Background

This project is conducted with the primary datasets provided by the Project Controls Unit at NYC Department of Design and Construction (DDC). The project's objective is to delve into the predictive analysis of construction project delays and their potential relation to cost in New York City using advanced machine learning and Natural Language Processing (NLP) techniques. Our methodology aligns with the DDC's project lifecycle as shown below, which includes distinct phases: Planning, Design, Bid, and Construction. Each phase represents critical junctures where predictive analytics can yield insights to mitigate risks of delays and manage project timelines. The construction phase, which is our project's focal point for risk management analysis, includes Pre-Construction, active Construction phases, Substantial Completion, Project Close-Out, and Final Completion stages. The predictive tool developed in Python will serve as an analytical instrument for the Construction Culture and Data Working Group and the Town+Gown program, providing them with a model to foresee and manage project delays. The final outcome of this project aims to enhance decision-making processes and improve the efficacy of construction project management within the city.



1.2 Problems Statement

The primary focus is the development of risk management tools. This tool aims to predict and quantify the risk of delays, drawing on historical data such as project phase, location, and narrative descriptions from the Project Control's archive of risks and delays. This will enable more informed decision-making and proactive mitigation strategies. While not the main focus, the secondary objective involves laying the groundwork for a predictive cost estimation tool.

1.3 Existing Work

Earlier studies of DDC construction projects have laid the groundwork for data-driven decision-making. However, the dataset underpinning our project is a novel compilation for a relatively short time period, which presents a unique opportunity as well as a challenge. The time limitation and new nature of the data mean that there is a paucity of directly relevant precedents or studies to draw upon. This gap underscores

the innovative aspect of our work, as it will contribute original insights and pave the way for future research in this area.

1.4 Overall Approach

We have access to three datasets for our project: the delay dataset, the portfolio dataset, and the cost dataset. Our approach to this project can be divided into 3 parts/stages: a thorough Exploratory Data Analysis to get a deeper understanding of the datasets, predictive modeling, and NLP techniques to understand the substantial description text data.

For EDA, we aim to visualize and analyze 3 datasets respectively. For the delay dataset, we want to understand aspects such as how delays are distributed among different boroughs, how delays are spread when grouped by individual project types, the visualization of the delay length distribution, and how delays vary over a period of time. For the portfolio dataset, we are interested in the connection between delay incidents and the type of sponsor agency that can serve as a proxy for project type. In terms of the cost dataset, we want to understand how labor, material, and equipment costs are distributed for every unit. Hence, we will plot and analyze each of these insights correspondingly.

With the goal of managing risks, delays, and costs, there are three predictive models that we want to build. First, the delay duration predictive model since delay duration is substantial in risk management and also explicitly relates to time where there could be associated increases in costs. Secondly, the delay category predictive model. This is designed to identify potential risks across different construction phases. Lastly, the delay prediction model. Here, we'll use the combined dataset to develop a predictive framework designed to forecast possible delays in new construction projects in New York City.

Finally, since there is a sufficient amount of text data involved in our dataset, we applied NLP techniques and analysis. This part is composed of 2 main aspects. Initially, they are model-led features. We examined the NLP-related features that had already been incorporated into our models. Subsequently, we explored deeply into the topic modeling process by which we utilized the descriptions of delays for two main reasons: firstly, to gain a more comprehensive understanding of the over 4,000 delays for NYC agency mentors; secondly, to reveal which factors mentioned in the text have an impact on risks, providing valuable insights for future studies on what comprehensive data to gather.

2 Dataset and Exploratory Data Analysis

2.1 Source and Overview

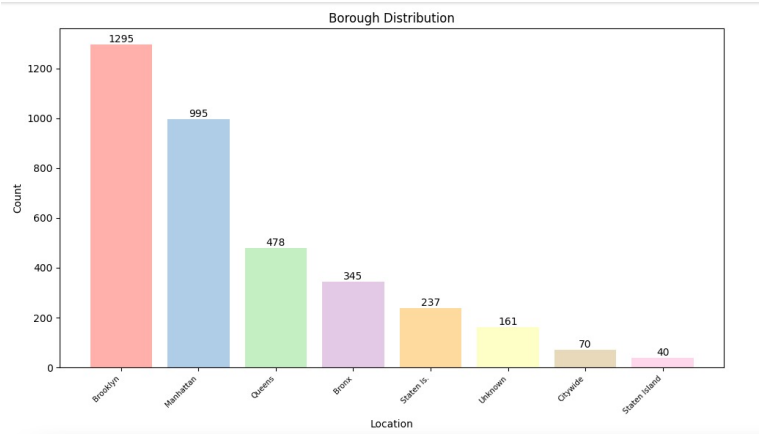
We have access to three datasets for our project: the delay dataset, the portfolio dataset, and the cost dataset. The delay dataset comprises 13 columns, predominantly featuring categorical variables. These include details such as the project's borough, type, delay start and end dates, and the phase of delay impact. A total of 754 unique projects are associated with delay text. Given that the number of unique project IDs is significantly less than the variety of delay descriptions, it's reasonable to infer that multiple projects experience more than one type of delay incident.

The portfolio dataset consists of 5,353 rows and 25 columns. The majority of these columns offer additional date-related information for each project, such as projected and actual start dates, as well as projected project closeout. This dataset also enriches our understanding by including categorical columns like benchmark status, project sponsor, and the current project phase.

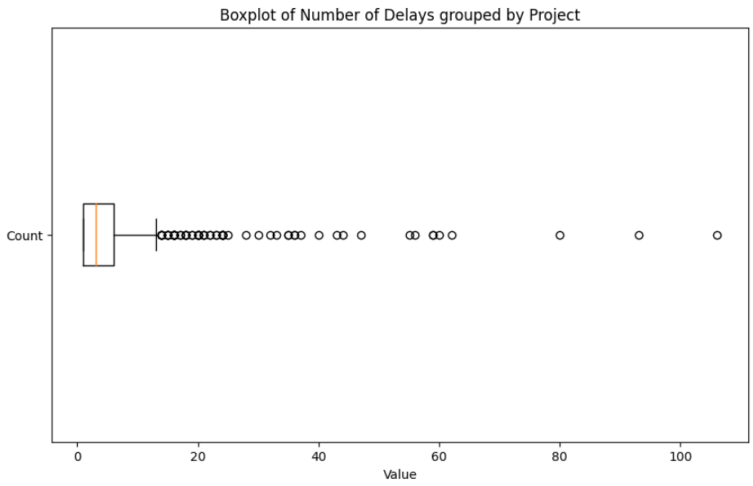
The cost dataset, comprising 42618 rows and 13 columns, presents detailed cost information for each construction project. Notably, the 'activity' column suggests that individual projects may have multiple activities, leading to potential duplication of project IDs. The dataset includes columns such as Labor UnitCost, Equipment UnitCost, Material UnitCost, etc. However, it's important to note that the cost columns are presented as string objects, which requires further preprocessing to convert them into numerical values.

2.2 Exploratory Data Analysis

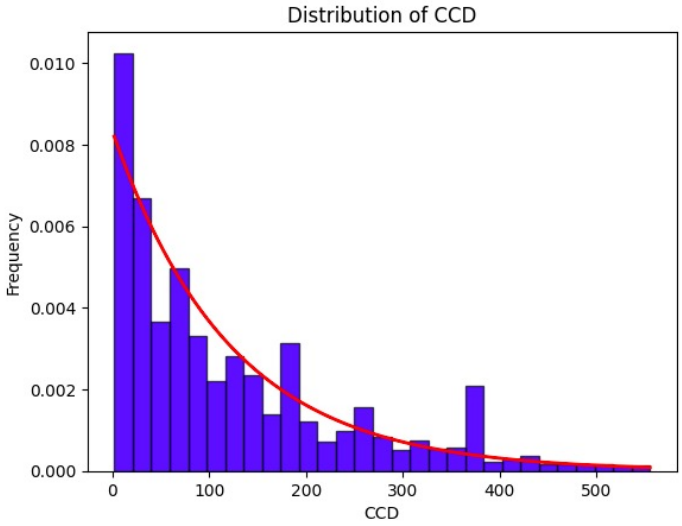
2.2.1 Delay Dataset EDA



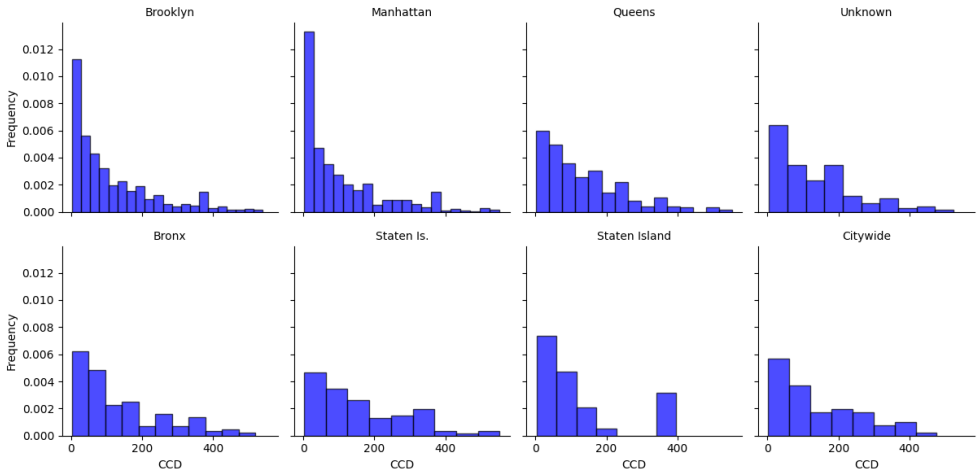
The bar plot depicting the total number of delay incidents across different boroughs in New York reveals insightful patterns. Notably, Brooklyn and Manhattan emerge as the focal points of delays, with frequencies of at least twice as high as the other boroughs. Considering that Brooklyn and Manhattan are the two most densely populated boroughs in New York, this observation suggests a potential positive correlation between the intensity of human activities and the occurrence of delays in construction projects. We applied methods similar to those for categorical variables other than borough, which yielded some interesting results. For delays across division distribution, 25% more delays occur for public building project types than for infrastructure project types. In terms of the delay category, we observed a major reason for the delay is construction conditions, which can be roughly understood as unexpected events or conditions during construction.



The boxplot provides a clear depiction of the distribution of delays grouped by projects. Most projects are impacted by a relatively moderate number of delays, typically falling within the range of 15 to 20 incidents. Only a few projects are impacted by more than 50 delays. We identified an outlier, a single project that experienced over 100 delays. This outlier adds a compelling dimension to the analysis, prompting further exploration into the factors contributing to such an unusually high number of delays in that specific project.

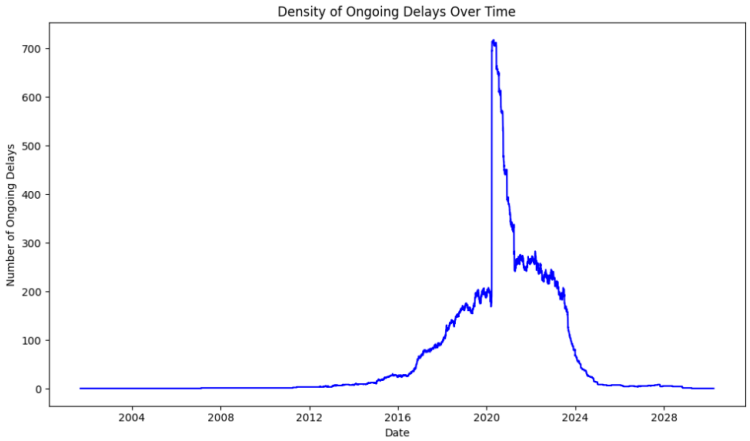


The histogram above shows the percentage distribution of all observed consecutive calendar days (CCDs). With the y-axis as the percentage (y-axis) and the x-axis as the number of days the delay lasts, we can observe that the percentage roughly follows an exponential decay. It's quite right skewed, indicating that most of the CCD values clustered around the lower end of the distribution, which also implies that the mean is largely influenced by the presence of higher values in the right tail.



We further applied several facet wraps, which yielded more exciting results. The above is subdivided by borough on CCDs. It is apparent that even divided into boroughs, almost each of them still roughly follows a right-skewed, exponential distribution.

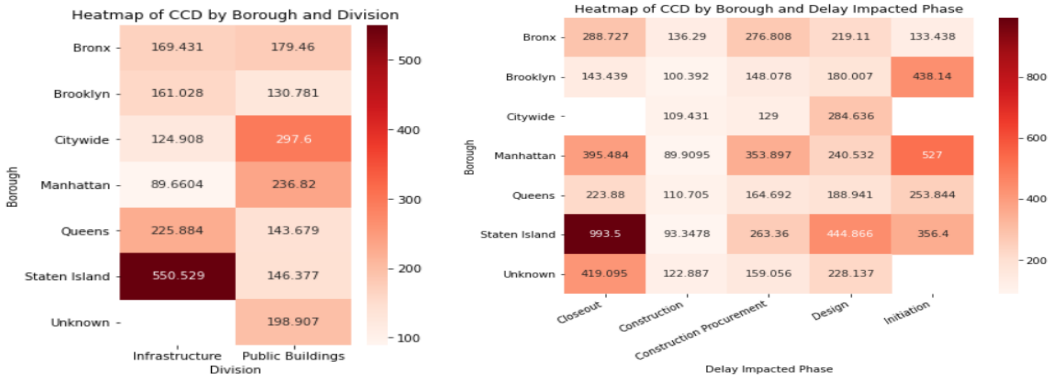
When it's subdivided by delay category, from the previous analysis that the two major categories of delays are construction conditions and external environments, we would expect that they should follow the overall trend since they constitute the central part of delays. However, although construction conditions follow our expectations, the external environment's impact follows a more linear rather than exponential delay. This might be the natural distribution of weather conditions in a linear decay manner. Similar phenomena also appear when we device the data by delaying the impacted phase.



Additionally, we are interested in the trend of the number of ongoing delays over time. The line plot indicates a relatively low and uniform number of delays from around 2004 to 2016, which might be due to a lack of data, records, or a naturally smaller number of projects considering the projects from the last decade. There is a gradual increase in delays from 2016 to 2020, potentially due to evolving project complexities or external factors. Then, we observed a huge spike, almost a vertical ramp, up around March 2020, which aligns with the time of the COVID-19 outbreak. From 2021 onward, the number of delays drops, possibly reflecting less strict lockdown policies. Starting in 2024, the number of delays decreased further down to the 2004-2016 level, with many being projected delays (future projects).

We then took a glance at an aggregation of multiple pairs of 2 categorical variables

on the mean of CCD through heatmap, which can offer an insight into which category has a specifically high delay.

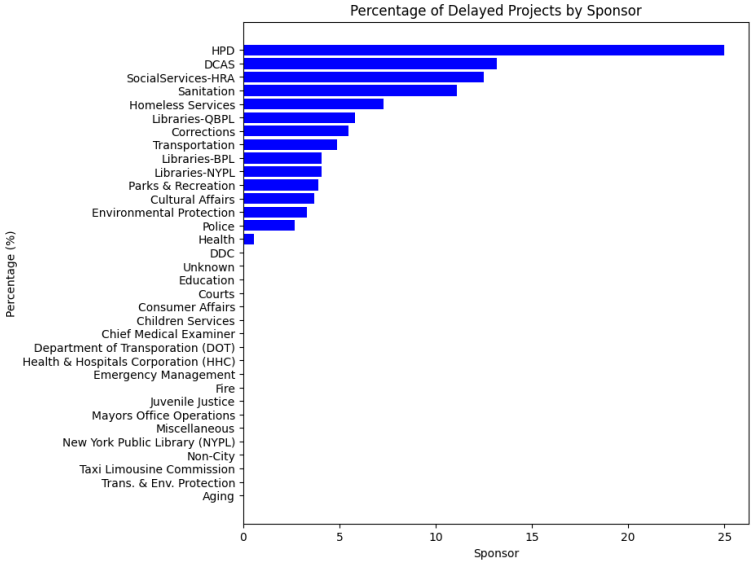


In the initial heatmap, we made an interesting observation regarding the distribution of CCD across different boroughs. Notably, the boroughs "Bronx" and "Brooklyn" exhibit relatively lower CCD values for both construction divisions, whereas all other boroughs display a notable imbalance. For instance, "Citywide" and "Manhattan" demonstrate higher average CCD values in the context of public building project types, while "Staten Island" stands out with an exceptionally elevated CCD in the context of infrastructure project type.

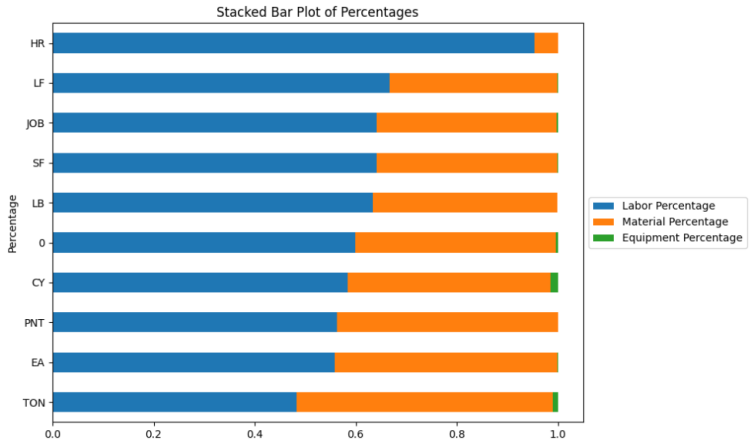
Moving on to the second heatmap, which pivots the data by borough and the phase impacted by delays, we uncover a consistent pattern. Across all boroughs, there is a general trend of lower CCD values during the construction phase compared to other project phases. Conversely, during the closeout phase, there is a marked increase in CCD, which aligns with our expectations. Notably, "Staten Island" particularly stands out with a significantly elevated CCD during the closeout phase.

2.2.2 Portfolio dataset EDA

In the portfolio dataset, we are curious to know whether there are connections between the delay incidents and each sponsoring agency as a proxy for a particular building or infrastructure project type. The barplot below depicts the percentage of delayed projects based on project types by sponsor agency. While several project types stand out for different proportions of delay, further analysis is necessary to understand these preliminary observations.



2.2.3 Cost Dataset EDA



In the cost dataset, every engineer's estimate (EE) unit cost is composed of labor cost, material cost, and equipment cost. This horizontally stacked bar plot is made in the interest of how these three parts are distributed for every different unit. We observed that the cost of equipment percentage is almost insignificant for all units. Labor cost is the major composition for every unit, with 'HR" having labor cost over 90 percent being the most remarkable one. For unit 'TON', labor cost is under 50 percent, which makes sense since 'TON' is a large, heavy-weight unit that doesn't require much human labor effort, so material cost is the most composition instead.

2.3 Preprocessing

During the data preprocessing phase of our study, we implemented standard procedures, including removing duplicates and imputing missing values, to maintain dataset integrity. We further refined our portfolio dataset, which contains extensive project information, by retaining only projects that had reached the close-out phase, as we could not predict future delays. This dataset was then merged with the delay dataset using Project IDs, resulting in a new binary column indicating the delay status of each project. Additionally, we excluded projects that began before January 1, 2020, since comprehensive delay data was only recorded starting from this date, ensuring our analysis was based on complete and accurate records.

3 Prediction Models on Risk Management

With the goal of managing risks and delays, there are three predictive models that we have implemented during our project: delay duration prediction model, delay category prediction model, and delay probability prediction model.

3.1 Delay Duration Prediction Model

3.1.1 Motivation and Introduction

The delay duration caused by risks is a critical aspect of risk management, as it directly relates to the time when the construction is underway, and delays can have cost impacts. Therefore, the first predictive model we implemented is the Delay Duration Predictive Model. Initially, we would like to predict the delay days to the exact day, such as 103 days, as a single numerical value. After that, we shifted our prediction objective from an exact day to predicting a range due to low accuracy.

3.1.2 Feature Engineering

For this model, we primarily utilized the DDC Project Control's delay dataset. It contains several directly usable features, such as project division (infrastructure or public building) and borough (Manhattan, Brooklyn, etc.). In addition to these features that we can directly incorporate into the feature engine, we also extracted more time-related and NLP-related features.

Based on the background knowledge of construction projects and corroborated by the time series graph we plotted, we observed a clear seasonality in delays. Therefore, we added the season as a new feature. Additionally, from our distribution graph, we noticed a significant peak in the number of delayed projects from the months following January 2021, suggesting that the pandemic period was one of the crucial factors for delays. As a result, we added a binary feature indicating whether the project occurred during the pandemic.

Next, we conducted an NLP-related analysis. There's a column in our datasets named delay description, which is a lengthy text. Based on this column, we undertook NLP analyses. We performed sentiment analysis, Term Frequency-Inverse Document Frequency (TF-IDF), and Latent Dirichlet Allocation (LDA) analysis, subsequently integrating these three features into the model. The following image shows an example of LDA analysis.

```
{0: ['main', 'con', 'gas', 'contractor', 'interference'],
 1: ['delay', 'consultant', 'design', 'due', 'work'],
 2: ['project', 'dep', 'plan', 'hold', 'moved'],
 3: ['due', 'covid19', 'order', 'emergency', 'governor'],
 4: ['due', 'delay', 'contractor', 'work', 'approval'],
 5: ['old', 'way', 'sheeting', 'need', 'make'],
 6: ['design', 'project', 'scope', 'ddc', 'due'],
 7: ['dep', 'additional', 'issue', 'work', 'requested']}
```

3.1.3 Model Methodology

We tried numerous mainstream machine learning models, such as ordinary least squares regression, random forest, etc.

At first, we tried to predict the delay days to the exact day, such as 103 days, as a single numerical value. At this stage, we primarily utilized regression models. We

employed ordinary least squares regression, random forest regression, and support vector regression. From the results, the best-performing model had an r-squared value of only 0.092, which is a highly unsatisfactory outcome. At this phase, we thought that the quantity and quality of our datasets were insufficient for our prediction model to be precise to the exact day: we had only about 4,000 rows of data, 735 projects, and seven columns that could serve as features.

After that, we shifted our prediction objective from an exact day to predicting a range. Based on historical data, we derived six quantiles: 0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, 0.8-0.95, and 0.95-1. Our goal then became predicting within which of these six quantiles a project's delay days would fall. As a result, we transitioned from using regression models to classification models: this became a six-class classification problem. We experimented with multi-nomial naive Bayes, K-nearest neighbor(KNN), support vector machine(SVM), random forests, and adaboost trees.

3.1.4 Results and Visualization

In the end, we retained the random forests model, which yielded the best performance, and did some optimization. Our optimization can be mainly split into two parts: tuning based on the model itself and tuning the feature engine. In the model-based tuning, we adjusted parameters such as maximum depth, number of estimators, and maximum features, ultimately selecting the best parameters. In the feature engine tuning, we introduced new features based on analysis and background knowledge, as described above.

At this stage, our six-class classification model achieved a final accuracy of 54.3%. The following chart shows precision, recall, and F1 scores for each classification quantile.

	precision	recall	f1-score
1	0.77	0.57	0.65
2	0.45	0.44	0.45
3	0.51	0.51	0.51
4	0.52	0.57	0.54
5	0.55	0.69	0.61
6	0.33	0.52	0.40

From the above chart, we can observe that the prediction precision for quantile 1 is the highest, at 77%. The recall score for quantile 5 is the highest, at 69%. Therefore, overall, the predictions for quartiles 1 and 5 are currently the best performing. This model is low-performed, and we think it's primarily because of limited data volume in terms of both project samples and features. Regarding this, we've done some posterior analysis and provided feedback to NYC agency mentors on future possible data synthesis strategies, which is illustrated in later sections.

3.2 Delay Category Prediction Model

3.2.1 Motivation and Introduction

This model identifies potential risks across different construction phases, such as in the sponsor and design phases. Each project could have several risks occurring during the whole construction. In varying stages, distinct types of risks can occur, leading to various degrees of delays and associated escalating costs. The dataset includes nine categories of delays, such as scope management and stakeholder management. Analysis from the EDA heatmap of delay categories reveals a distinct distribution of delay

types across each construction phase. This observation is corroborated by background information from NYC agency mentors, confirming the fact that each phase is characterized by different risks with varying probabilities. For instance, scope management issues typically arise in the earliest stages. Consequently, after consultations with the mentors, developing a model for predicting risk categories would be essential to effective risk management.

3.2.2 Feature Engineering

In addition to the newly extracted time-related features of delays in the previous model, such as the season when the delay occurs, we have also merged new features from the data portfolio dataset. As mentioned in the model introduction, each construction phase has different probabilities for various delay categories. Therefore, phase-related factors have been incorporated into the engine, such as the month and season when the design phase and construction phase begin. Additionally, insights from the risk category and background information indicate that certain risks, like stakeholder management, can also probably be associated with specific sponsor agencies as proxies for the project type. Consequently, the sponsor column from the data portfolio dataset, including program unit, activity, and other background-related features, has been merged into the engine for different risks.

3.2.3 Model Methodology

This model remains a classification model. We initially attempted multinomial logistic regression, as theoretically, with limited data volume in terms of both projects and features, a simpler model might yield better accuracy. We also tried mainstream machine learning models, including SVM, random forests, and AdaBoost trees. However, the final outcome still showed that the random forest yielded the highest accuracy.

3.2.3.1 Model Tuning and Feature Selection

We commenced model tuning, with feature selection being the most crucial aspect. As mentioned earlier, features like sponsor agencies as proxies for the project type, each with over 50 unique values, resulted in more than 50 dummy variables in the model due to their categorical nature. The feature importance analysis from the random forest indicated that some dummy variables had very low importance values, with some even at zero.

We employed three methods to identify important features. The first was using a correlation matrix, but perhaps due to the abundance of features (total 145 features including many dummy variables), we couldn't find any feature with significantly high or noticeable correlation values with the risk category.

The second and third methods involved selection based on the importance of random forest features. One approach was a straightforward selection of the top N important features. With 145 features in total, including many dummy features, we adjusted the value of N to select the top N features and then reintegrated them into the model to observe any improvement in accuracy. This approach also involved choosing features exceeding a certain importance value threshold. The other method added an additional step of considering feature dimension stratification, selecting the top N dummy features within each feature category or those exceeding a certain importance value.

3.2.3.2 Possible Relationship within Risk Categories and Correlation Algorithms

Prior to model implementation, we also contemplated the possibility of interrelationships among the risks, such as the occurrence of one risk potentially increasing the likelihood of another, given that multiple delay categories often occur within a single project. After consulting with our mentors, we agreed this was a plausible scenario. To verify this hypothesis, we employed correlation-related algorithms. The following figure presents some of the interrelationship information obtained using the Apriori algorithm.

	antecedents	consequents	antecedent support	consequent support	support
0	(2 Schedule Management)	(8 External Environment)	0.301061	0.645889	0.201592
1	(3 Cost Management)	(8 External Environment)	0.152520	0.645889	0.106101
2	(4 Quality Management)	(8 External Environment)	0.088859	0.645889	0.075597
3	(5 Resource Management)	(8 External Environment)	0.140584	0.645889	0.103448
4	(6 Procurement Management)	(8 External Environment)	0.224138	0.645889	0.147215
5	(7 Stakeholder Management)	(8 External Environment)	0.124668	0.645889	0.091512
6	(9 Construction Conditions)	(8 External Environment)	0.251989	0.645889	0.155172
7	(1 Scope Management, 2 Schedule Management)	(8 External Environment)	0.118037	0.645889	0.090186
8	(1 Scope Management, 3 Cost Management)	(8 External Environment)	0.063660	0.645889	0.053050
9	(1 Scope Management, 4 Quality Management)	(8 External Environment)	0.037135	0.645889	0.030504

Antecedents: These are the precursor delay categories in our projects. Consequents: The delay categories that often follow the antecedents. Support: Indicates how often the antecedent-consequent combination appears in our dataset.

The chart displays the top 9 antecedent-consequent pairs with the highest support values, revealing that all 9 consequents are "External Environment." The data also indicates that the "External Environment" is the most common type of delay, comprising 64.5% of cases. Therefore, its high frequency may erroneously appear to strengthen its association with other risks. Nevertheless, the outputs show that 8 antecedent-consequent pairs have a support value greater than 7%. Given that there are just over 700 projects, a 7% probability is significant enough to warrant analysis. Furthermore, 5 pairs exceed 10%, with two pairs surpassing 15%.

Based on the Apriori algorithm and our mentors' expert understanding, we believe that risk interrelationships do exist among certain risk categories. Consequently, these 8 antecedents can become important features in the model for predicting the likelihood of the consequent risks occurring.

3.2.4 Results and Visualization

The following graph displays the overall accuracy of the model as well as the predictive accuracy and recall scores for each of the nine risk categories.

	precision	recall	f1-score	support
1 Scope Management	0.39	0.37	0.38	90
2 Schedule Management	0.33	0.39	0.36	93
3 Cost Management	0.30	0.28	0.29	46
4 Quality Management	0.21	0.24	0.23	25
5 Resource Management	0.22	0.17	0.19	40
6 Procurement Management	0.22	0.22	0.22	60
7 Stakeholder Management	0.13	0.20	0.16	25
8 External Environment	0.38	0.43	0.41	171
9 Construction Conditions	0.85	0.75	0.79	331
accuracy			0.49	881
macro avg	0.34	0.34	0.34	881
weighted avg	0.52	0.49	0.50	881

The weighted accuracy of the model is 52%. Given the limited data volume, this result has already surpassed our and our mentors' expectations. It is evident that there is a significant variance in the accuracy of risk classification. The precision for construction conditions risk reached 85%, with a recall of 75%, while the accuracy for other risks

was as low as 13%. We attribute this to the composition of the dataset's columns.

Many inputs in the columns are related to construction conditions, whereas, for risks with lower accuracy, such as stakeholder management, the related inputs might only be the sponsor's name. However, as observed in the correlation matrix, the presence of certain sponsors does not significantly indicate this type of risk. Moreover, with over 50 different sponsors, the data becomes more dispersed, making it challenging to validate this conclusion. With other risk categories, for example, resource management, there are no direct inputs related to it. From the EDA and background information, only some indirectly related factors, such as the borough location and activity, exist. Hence, in the absence of these direct and effective inputs, the model shows poor performance in accuracy and recall for these risks.

Based on this observation, we contemplated a potential future roadmap: with professional knowledge of these risks in the industry, can we integrate more comprehensive datasets, combining data from various groups? For example, resource-related data are primarily managed by a group in the department, or integrating data from the stakeholder side, such as possible reasons for their induced delays or their resources information, could be beneficial. Integrating data from various groups would undoubtedly significantly improve the risk prediction model.

3.3 Delay Probability Prediction Model

3.3.1 Motivation and Introduction

For this part of our project, we aimed to establish a predictive framework that anticipates potential delays in new construction projects within New York City. Utilizing the integrated dataset that merges portfolio and delay data, we can gain insights into patterns and correlations that might signal impending project setbacks. The dataset initially comprised three categorical predictive variables: division, program unit, and sponsor. We also enriched it with derived features to encapsulate the circumstances of the COVID-19 pandemic. The primary goal was to leverage this enriched dataset to forecast delays in the construction phase. Given the current limitations of our dataset, our model serves as an initial step towards a more comprehensive predictive system. It is designed to evolve and become more refined as additional data is integrated, embodying a dynamic tool that adapts to new information and improved methodologies.

3.3.2 Feature Engineering

In response to the substantial impact of the COVID-19 lockdown observed during our EDA, we introduced two new binary variables to our project dataset: 'ended_before_lockdown' and 'started_after_lockdown.' These variables are self-explanatory by their names. 'ended_before_lockdown' indicates whether a project reached its close-out phase prior to the onset of the COVID-19 lockdown in New York City, whereas 'started_after_lockdown' denotes whether a project commenced subsequent to the initiation of the lockdown measures. The inclusion of these variables was instrumental in quantifying the lockdown's influence on project timelines, allowing for a more nuanced analysis of pandemic-related disruptions.

3.3.3 Model Methodology

Prior to delving into the methodology of our predictive classification model, it is imperative to acknowledge a notable limitation regarding model tuning. The merged

dataset at our disposal comprises merely 97 entries, a figure insufficient to capture the broader spectrum of scenarios. This limitation in data volume may potentially introduce significant bias, diminishing the model's ability to generalize effectively. Consequently, the insights derived from this model should be interpreted as preliminary guidelines and references. Future refinements and enhancements to the model will likely be necessary as more comprehensive datasets become available.

As mentioned before, this model aims to predict whether a project will be delayed or not during the construction phase according to the categorical information we acquire as the project is initiated. In our pursuit of an optimal predictive classification model, we experimented with various algorithms and methodologies. Ultimately, we selected the Catboost model, renowned for its effectiveness in handling categorical data and its robust performance with limited datasets. To address the challenge posed by our dataset's imbalance, we also employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is particularly beneficial in enhancing minority class representation, thereby mitigating biases towards the majority class. However, it is essential to note that due to the limited size of our data, we were unable to substantiate our choice quantitatively with extensive empirical results since the results are easily overfitted or biased. Theoretical advantages of the Catboost model, such as its built-in handling of categorical variables and gradient boosting approach, combined with SMOTE's proven efficacy in addressing class imbalance, formed the basis of our decision. As more data becomes available, we anticipate further validation and potential adjustments to our approach, aligning with the evolving dataset characteristics.

3.3.4 Results and Visualization

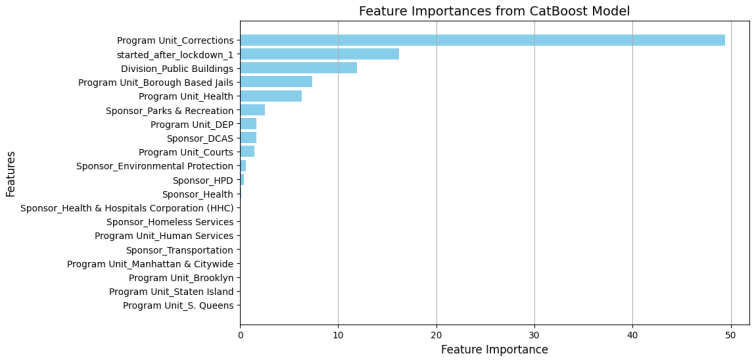
The following result shows the test set metrics of our constructed pipeline. As we can observe, the test metrics reveal that our Catboost model exhibits a high level of accuracy at 85%, reflecting a solid predictive performance for the negative class. The F1-score for the positive class, however, suggests an opportunity for model enhancement, as it currently stands at 0.00. This highlights the model's potential for further refinement, especially in the context of our constrained dataset. We anticipate that with a larger dataset, the model's ability to discern across different classes will improve, leading to more balanced and representative performance metrics.

```
Test Set Metrics:
Accuracy: 0.850
F1-Score: 0.000
              precision    recall  f1-score   support

           0       0.85      1.00      0.92        17
           1       0.00      0.00      0.00         3

    accuracy                           0.85        20
   macro avg       0.42      0.50      0.46        20
weighted avg       0.72      0.85      0.78        20
```

We also produced the feature importance plot of the Catboost Model. We can observe that the Program Unit_Corrections emerges as the most influential feature, holding the highest importance score and potentially playing a pivotal role in the model's predictive power. The feature with the second highest importance is 'started_after_lockdown,' which aligns with our understanding of the COVID-19 lockdown's impact on project timelines. There are other features, such as the building or infrastructure typology, that suggest considerable importance and require additional analysis and more data that becomes available to predict outcomes.



In conclusion, our Catboost model establishes a preliminary pipeline for project outcome prediction. It's important to note that the model's current parameters and identified feature importances are initial insights that are expected to be refined with additional data. Our goal is to provide a scalable framework that will enhance accuracy and reliability as further data becomes available for analysis.

4 NLP Analysis and Topic Modelling

In our dataset, there is a substantial amount of text data about delay description, which makes NLP analysis a focal area of our study. The NLP-related aspects can be primarily divided into two parts: one for predictive model features and the other for topic modeling.

Firstly, we will discuss which NLP-related features have already been employed in our model. Secondly, we delve into topic modeling for two primary reasons: firstly, to help with mentors' need to get a comprehensive and summarized understanding of delay incidents; secondly, as mentioned in the model section, there is a future need to amalgamate data from various groups. Thus, we can utilize the descriptions of delays for posterior analysis to assist in identifying which data are imperative to be integrated.

4.1 Model features

In the modeling section, we mentioned extracting certain information from the context data to serve as features for our model. This primarily involves sentiment analysis, TF-IDF, and LDA analysis. We incorporated sentiment as a feature for predicting the duration of delays, hypothesizing that more negative sentiments might lead to longer delays. TF-IDF and LDA analysis can be understood as clustering features. This involves categorizing projects and delays into N topics to assist in predicting the duration.

From the accuracy and feature importance of models, it is evident that these NLP-related data have contributed to an improvement in performance, although the increase is not significantly high. Regarding this, we speculate that the existing NLP packages may not sufficiently align with the specific context of construction in New York. Thus, a potential future direction could involve building upon current NLP packages and using deep learning algorithms, allowing fine-tuning according to our specific construction scenarios and dataset. We had considered attempting this, but the current text data pertains to delay descriptions, which are summaries made after a delay has occurred. Therefore, using this to train the model is ineffective, as these descriptions are conclusions and not information we can obtain for new projects (except in the risk duration model, since even after a delay's occurrence, its exact duration remains uncertain, which is why we only used delay description-related NLP features in the first risk

duration model).

However with delay descriptions, we can still utilize them for posterior analysis to help future data collection, which will be elaborated in the next section.

As mentioned in the model section, if we can obtain comprehensive preliminary data from various groups, such as preset descriptions about project type, stakeholders, resources, quality, etc., for new projects, a fine-tuning process based on the current NLP predictive model could be highly beneficial. For instance, we could use Bidirectional Encoder Representations from Transformers (BERT) as a base word encoder model since it's potentially the best base for fine-tuning classification tasks. On the head layer of BERT, we could build a supervised classification neural layer tailored to risks, helping to align extracted topics better with delay categories or probabilities. Alternatively, we could also train the model so that text embeddings of the same risk type have closer cosine similarity. We have also provided feedback on these aspects, and our NYC agency mentors have adopted our suggestions regarding data integration.

4.2 Topic modeling

This section is primarily focused on topic modeling. Through topic modeling, we can identify prevalent themes and topics within the delay descriptions, which can highlight recurring issues or factors contributing to delays. The rationale behind these analyses stems firstly from gaining a more comprehensive understanding of the over 4,000 delays for NYC agency mentors. Secondly, it's also a posterior analysis that helps inform future data collection strategies. By recognizing the influential factors of delay mentioned in these text data, mentors and groups can tailor their data collection efforts to capture more relevant and detailed information, potentially improving risk prediction and management in future models and projects.

4.2.1 Entity Extraction

We initiated our analysis with the named entity extraction. The following image displays the frequency of some entities and examples in our dataset.

Entity Type	Frequency	Examples
ORG	4549	['FEP', 'SD', 'MAS']
GPE	221	['SF', 'BQE', 'Kensington']
FAC	140	['City Hall', 'Leonard Street', 'Richardson Street']
ORDINAL	128	['44th', '45th', '46th']
PRODUCT	119	['Task Order #5', 'the Bulletin 1', 'the Bulletin 1']
QUANTITY	92	['2-foot', '1901 six inch', 'only eight inches']
LOC	84	['existing72', 'the Cost Estimate', 'Estimate']
NORP	59	['Design', 'M-5510', 'Jewish']
WORK_OF_ART	50	['TWM', 'SM', 'TWM']
LAW	46	['chamber No.2', 'Gas Services', 'STO']
TIME	25	['minutes 7/21/22', 'working hours', 'daytime hours']
EVENT	14	['Phase II', 'Phase II', 'Federal FY 2021']

These entities, as mentioned earlier, assist in providing a more comprehensive understanding of delays. For example, we can identify high-frequency locations where risks occur and laws that impact construction.

For future projects, this allows us to purposefully observe and collect information related to high-frequency entities. For example, we can see that "ORG" (Organization), "GPE" (Geopolitical Entity), and "FAC" (facility) entity have relatively high frequency, for new projects,is that possible to collect data on these entities and put them as features in prediction models.

4.2.2 N-gram frequency

This part of NLP analysis involves n-grams, which can be understood as extracting and counting high-frequency words or phrases.

A critical aspect of this analysis is word cleaning, which is necessary due to individual word usage habits or contexts. This involves custom mapping singular and plural forms of the same word together, lemmatizing, and mapping nouns. and verb forms, artificially removing high-frequency yet insignificant words, such as 'due', 'work', and some custom stop words, etc.. Removing words is particularly relevant for 2 and 3-gram analysis, where we examine the word preceding or following a target word because it can make adjacent words more meaningful and relevant.

We began with a 1-gram frequency analysis. The following image illustrates a word cloud of single high-frequency words.

We retained words with a frequency greater than 50 after cleaning, a total of 195 words. After discussion with mentors, we keep around 30 words. Regarding these 30 words, we further extracted 2 and 3-gram frequencies to gain a better understanding of delays. The following image shows an example of 2 and 3-gram frequency extraction.

omb 2-gram		3-gram	
Words preceding 'omb':		words preceding 'omb':	Words following 'omb':
delay: 11		cp send: 7	review approval: 5
send: 10		bid contract: 4	cp review: 5
approval: 8		cp approval: 4	review ewa: 4
submit: 8		cp submit: 4	prolong evaluation: 4
review: 7		request submission: 4	proceed directive: 4
submission: 7		hold cp: 4	8 12: 4
cp: 5		fund transfer: 3	request cp: 4
pende: 4		project meet: 3	target budget: 3
contract: 4		submit submit: 3	sufficient funding: 3
19: 4		sponsor agency: 3	inquiry cp: 3
		review approval: 3	4 register: 3
Words following 'omb':		review submission: 3	lack staffing: 3
review: 15		cp review: 3	multiple eligibility: 3
approval: 13		advertising authorization: 3	design register: 2

Similar as described before, it's also helpful both for a comprehensive understanding of delays and future data collection, for example, "utility" is highly frequent, so can we have a specific column about utility for future projects?

4.2.3 Feature distribution after N-gram frequency

Finally, based on the 2 and 3-gram frequencies, we further narrowed down to about 10 high-frequency phrases that are significant for risk analysis. We conducted an EDA-like analysis: when these phrases appear, what is the distribution of other features of the project? For instance, what is the project division distribution when an 'emergency order' occurs? Is 'utility interference' more likely to occur in certain construction phases? Understanding these distributions is immensely helpful in comprehending the risks involved.

	Public Buildings	Infrastructure
change order	42	12
order registration	32	6
direct interference	0	107
interference propose	0	1
utility interference	0	44
con edison	6	412
request sponsor	19	1
sponsor request	19	1
field condition	6	10

The above table is like the heatmap of the distribution of "Division" features across these keywords. We observed that 'direct interference', 'utility interference', and 'con edison' appear mostly in infrastructure. Notably, 'con edison' appeared 412 times. Considering Con Edison is an electricity, gas, and steam provider to NYC, this indicates that during infrastructure construction, utility-caused delays are common.

We got similar tables for the other features, like "Delay Impacted Phase", and "Borough". For example, for "Borough," we also observe that "utility" and "interference" are frequent for "Manhattan" and "Brooklyn".

Therefore, after these results, we can indeed consider collecting specific columns about utility and interference, or even if the project highly relies on "con edison" or not.

A	B	C	D	E	F	G	H	I	J
	EXTERIOR RENOVATION	MISC	(B) Major Rehabilitation/ Renovation	INTERIOR RENOVATION		(A) New Construction	Sewer	Green Infrastructure	Highway
change order	8	2	9	18	12	5	0	0	0
order registration	10	2	2	16	6	2	0	0	0
direct interference	0	0	0	0	107	0	0	0	0
interference propose	0	0	0	0	1	0	0	0	0
utility interference	0	0	0	0	44	0	0	0	0
con edison	1	1	2	1	412	1	0	0	0
request sponsor	9	0	6	4	1	0	0	0	0
sponsor request	9	0	6	4	1	0	0	0	0
field condition	1	1	1	3	10	0	0	0	0

The above table is the 'parent project type' feature distribution across these keywords. We observed that a significant proportion of these keywords fall into the null column, which suggests that many of these projects lack a parent project type. Therefore, we can try to improve this parent project-type data for future projects. Moreover, according to the background information, issues that affect infrastructure projects, as shown in the first table, should also affect public buildings. However, it's interesting that they didn't occur for sewer, highway, and green infrastructure, where we would expect them to. Thus, specific to these three parent project types, we can conduct further analysis to discover which factors cause their delays so that it can be helpful for future projects in these categories.

5 Summary

5.1 Conclusion

In conclusion, our project with the NYC Department of Design and Construction marks an initial foray into harnessing machine learning and NLP to enhance construction project oversight. Despite the dataset's current constraints, the application of these analytical tools has provided a valuable and predictive perspective on project delays from multiple aspects. The incorporation of NLP has been particularly insightful, providing a deeper understanding of the textual data associated with delays, enriching the predictive model, and contributing to a more informed risk management strategy within the existing framework.

5.2 Future Roadmap

The main future roadmap is to continue NLP analysis on delay descriptions to identify factors influencing delays mentioned in the text, facilitating future data collec-

tion and integration. After the final presentation next week, we will follow up on this process with mentors. The current plan includes:

As mentioned earlier, many expected terms related to sewer, highway, and green infrastructure did not appear. We will continue researching to identify factors impacting them. Analyzing the distribution of project types (50+ project types) when key phrases occur. Probably incorporate two additional steps: "topic coherence" and "dependency parsing" to achieve a better understanding of delay summaries.

5.3 Ethical considerations

In the realm of utilizing New York City administrative construction data for our data science project, identifying and addressing potential ethical issues is imperative. One key concern is the safeguarding of privacy and confidentiality, particularly in handling sensitive construction information. To mitigate biases within the dataset, we implemented thorough checks and corrections during the analysis process. Additionally, obtaining explicit consent for data usage is a crucial step in ensuring ethical practices. By incorporating these measures into our approach, we proactively work towards resolving and avoiding potential ethical pitfalls associated with the use of this construction dataset.