Developing Predictive Model for Project Cost Estimate

Scope: NYC DDC has initiated a machine learning project to develop predictive model for estimating cost of project and work items. Using the latest technique in Machine Learning and Advanced Statistics, NYC DDC to develop a model that predicts the cost of future and active projects and construction work items in different phases of the lifecycle of the project based on historical data. DDC has partnered with Microsoft who is providing the proof of concept guidance and making tools available for the proof of concept development. DDC is seeking assistance of a data scientist from the Town and Gown program to develop the model. A fully developed model is expected to help the user answer following questions.

System (Project Type) Level

- If an order of magnitude cost estimate is needed for a conceptual project, how can we identify 10 comparable projects that are most similar to the project(s) at hand, as a benchmark?
- Can we forecast parametric cost for different systems– For example how much does a major renovation of a library of certain capacity cost? How much does a window replacement project cost per window?

Project Level

- We have identified 20+ project complexity factors that influence project cost to different extents. Can we adjust cost of a project based on the combinations of applicable project complexity features such as local urban density, congestion levels, building age, project type, sponsor, duration, etc.?

Activity Level

- Can we forecast costs of certain type of repeat construction work items such as stair case demolition, building elevator installation per floor, foundation per cubic foot, copper pipe insulations per liner foot, brick removal per cubic feet, etc.?

Bid Analysis

- Is there a correlation between bidding configuration (number of bids, identity of bidders, bidding period, type of work) and market conditions and the project cost?
- Can we identify work items that have highest variance between bid and DDC estimates? e.g.) 'Hoisting / Distribution / Elevator time' work type of a general contract drives 50% of the contract price, etc.

Dataset: With 250+ projects in the dataset has detailed data points for project unit cost and project size. Each project has hundreds of construction work line items with their respective costs. The dataset includes basic project information as well as project type and contract type as well as complexity factors that specifically represent certain characteristics of projects. The data is cleaned and transformed for Power BI dashboard development – it can be used for initial exploratory analysis

Available Resources and tools:

- DDC Project Controls team for knowledge in construction project management
- DDC Data Analytics team for data description, generation of data points through process as well as fundamental statistical analysis of the data
- Microsoft AI team for technical advice as well as advice for Azure analytics tools
- Microsoft ML platform
- Microsoft Power BI

Tasks expected from the data scientist candidate:

- Explore and understand the data provided by DDC

- Communicate iteratively with DDC Data Analytics team and Microsoft AI team to discuss meaningful insights
- Interpret the data and draw meaningful inferences especially dealing with small datasets
- Test and select a most optimum algorithm to build a model

Desirable Technical Skills:

- Advanced statistical analysis and drawing inferences when dealing with relatively small datasets
- Solid understanding of data distribution analysis and hypothesis testing
- Experience in performing analysis with various Machine Learning algorithms
- Microsoft AI team will provide guidance on MS Azure ML, proof of concept, techniques in Machine Learning algorithms and usage of Azure ML functionalities.