Berkeley Data Analytics Final Project:
Predictive Model for Building Project Budgeting

Loc Nguyen

Ashok Ramaswami

Background

- We are working on data from a company that provides energy optimization and related services for buildings of various types. (One of our team members works for the company.)
- Most of the company's work is project-based. The company bids for these projects and delivers services according to the scope of the client's requirements.
- The company has been successful in their work but is interested in exploring if their data can help them make better pricing and planning decisions?

The question for our project: Can we estimate whether a given project is likely to be <u>over or under budget</u> based on the bid and historical patterns?

Data Source and Description

- **The company has detailed data for 1300+ projects, including the following features:**

  Project team
  Project manager
  Service line
  Building region
  Budget
  NDA
  Total hours
  Project duration
  Employee count
  Over budget or not

Analysis Objectives

We have the following broad objectives for our analysis:

- Build a model that predicts the likelihood of a project being under or over budget
- Test the model with different features and ML models
- Recommend areas for further improvement and model development

Description of Data Exploration Phase

- Dataset contains 4 tables: Building, project, hours, and employees.
- As a first step we removed any columns that we do not use (because they are descriptors and not used for prediction)
- Second, we cleaned up rows to exclude any null/NaN data.
- Our target label, "overbudget_yes", should be boolean, so we deleted any records with value "No conclusion!"
- Finally, we converted data types, mostly for the date columns, and renamed columns for easier tracking and analysis. After that, all tables are ready to import to our SQL database.

Creating our SQL Database

- We chose SQL to create a database with our tables and generate the dataset for our machine learning model
- For the SQL database:
  - Merged the project table with the building table in order to extract the building information for each project,
  - Merged the hours table with employees table to get the employee region for each hour entry.
- An exploration of the data showed that many projects are executed by either only one employee, or only in one region. Therefore, using the employee region proportion feature might make the model overfit, so we decided to remove this feature.
- Our final list of the features are: service line, project tema, project manager, budget, building region location, project having a NDA (non-disclosure agreement), total number of hours spent on the project, total number of employee working on the project, and project month duration.

Description of Analysis Phase

- Though the data we used is fairly clean, we still needed to preprocess the data before creating a model.
- The cleanup process included:
  Summing the number of hours from the hour table to the project
  Calculating the project month duration using the first date/last date of the hours table and the project end date
  Counting the number of employees working on each project.
- Following this cleanup process, we prepared the dataset for one-hot-coding, scaling, and generating a model.

ML model

- We created a dataset based on the feature engineering that we described in earlier slides
- Specifically we created ML models with two approaches:
  - SVM
  - Random Forest
- We applied the standard form for separating training and test data
- We then reviewed the model accuracy for each approach
- In addition, we ran a cross validation to evaluate model accuracy under different test samples

Dashboard Visualization

- Our goal is to create a simple dashboard that would take some user input parameters (such as project budget, duration, number of employees, estimated project manager, estimated hours, etc.) and predict the target
- We hope to refine this plan further based on input from instructors and assessment of feasibility
- We'll document any updates in our future submissions