Avi Kumar Talaviya Project Puzzle Report

Major Predictions and Approach

What trends or patterns influence project delays and cost overruns?
A: First of all, datasets are really complex to analyse, and because there isn't any data dictionary as well as the relevant context of these datasets.

- I analysed the Costrac dataset to understand the various cost involved in the execution of project
- In my analysis, I found that around 8 features affect the cost overrun namely `['CostType', 'ElementCode', 'DisciplineCode', 'PackageCode', 'Item',' Contractor', 'CostClass', 'TransactionType']`
- These features are influencing the pending or approved changes in the budget in a positive or negative manner
- While Original Budget values are missing in the dataset, I assumed only to look at the changes in budget which are mentioned in 'PendingChanges' and 'ApprovedChanges'
- these features include mainly either items/materials used in projects or contractors who executed these projects. Either of them can cause cost overruns or changes

What is the reasoning behind your answer?
A: In my analysis I found that features highlighted with yellow marks are ones that belong to either the input item to the project or the contractor who built these projects

- For example, Items with ElementCode 'AE' and 'CCS' has some positive pending changes while other ElementCodes have approved changes in budget lines
- Similary, CostType, and DisciplineCode has some pending changes or approved changes which means that these items are having the cost overruns
- Same with TransactionType and CostClass features have some pending or approved changes in their budget
- Interestingly. Those items or rows that have a positive change in cost have missing contract names.
- Not all categories of each feature are in positive change in budget because items which have positive changes are about 25% of the dataset while about 26% of the dataset has negative changes in budget

How robust are these patterns, and how might they be improved?
A: Patterns found in this dataset were significant and there are some categories that have positive changes in either pending or approved budget

Which datasets enabled you to understand patterns relating to the key questions for the project?
A: I could analyze the Costrac dataset for this project. Other datasets lack information about feature descriptions and some background info on how they have recorded the values of the datasets.

What data gaps have you identified?
A: Data descriptions were missing for each dataset.

What additional data do you believe is required to enable you to make predictions?]
A: We can make predictions on cost overrun from the ‘Costract’ dataset but I can’t say about the project schedule

Key Features or Metrics

What are the key features or signatures in the data that help us predict project delays and cost overruns?
A: `['CostType', 'ElementCode', 'DisciplineCode', 'PackageCode', 'Item',' Contractor', 'CostClass', 'TransactionType']`

- These features help in predicting cost overruns of the project but in all these contractor features have many missing values.

Did you use any additional external data?
A: No, I did not use any external data

Did this data improve your ability to predict project delays or cost overruns?
A: This dataset helps to predict cost overruns but difficult to conclude anything about project scheduling and delays due to a lot of missing values as well as a lack of further info on datasets

Methods and Approach

What is the best analytic approach to take?
A: To analyze such data, It is a good approach to divide the dataset into two parts: 1) Items/Input costs 2) Contractor features

- By diving the dataset into parts, we can understand how a pre-construction phase causes a cost overrun as well as the construction phase
- Each feature has a few of categories that have some pending or approved changes in budget
- We can also analyze all eight features together with cost-related features

If you think machine learning is viable, what type of algorithm would you use?
What did you try?
A: We can build ML models to predict if a certain item will result in a cost overrun or not. And also we can make a regression model to predict by how much amount it can run into cost overrun.

- Models like Decision trees, random forests, gradient boosting, and Naïve Bayes can be useful for such tasks. I haven’t tried on these models as the main purpose was to analyze data and find insights about the dataset.

Did anything not work?
A: I am finding it difficult to tackle and analyse GRC and Project scheduling data

Data Preparation

How did you prepare the dataset for analysis?
A: I, first of all, loaded the dataset and try to comprehend metadata as well various features of the dataset. I followed the below steps for data prep

- Costrac dataset has most of the features without missing values. Only the ‘Contractor’ feature has many missing values but this feature can be ignored for analysis (i.e removed). We can add any value to this feature because these are the names of the contractor
- Then, I created sub-data frames of items/input products and contractor-related features for further analysis.

What transformations did you perform?
A: I created a sub-data frame to analyze both datasets independently.
Once Analysed both sub-datasets I joined 8 influencing features with cost-related variables.
Where I created 2 new features for analysis

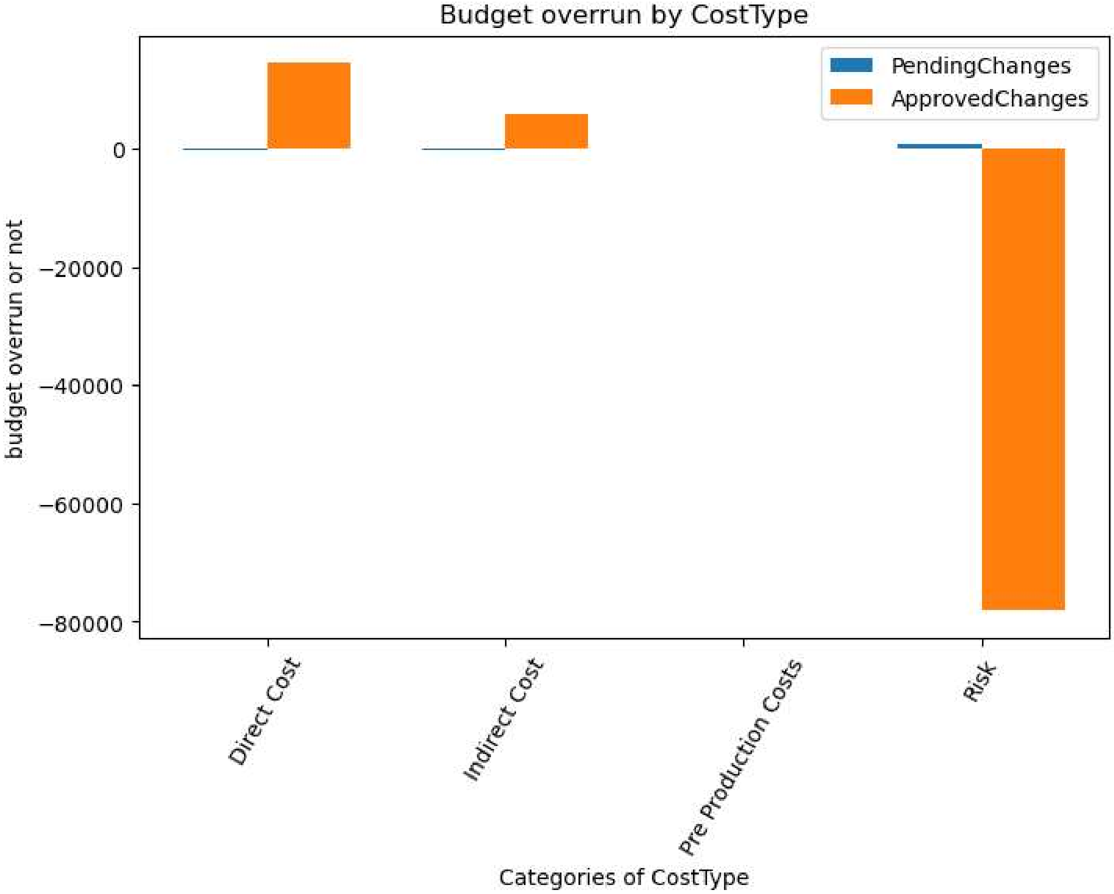
How would you recommend dealing with data like this?
A: Such datasets required an extensive understanding of background information about how and from where these datasets are taken.

- I would say transform such datasets into smaller chunks to understand each feature better. Also, other way could be analyze data from each project wide independently rather than a joint dataset of all tasks and features.

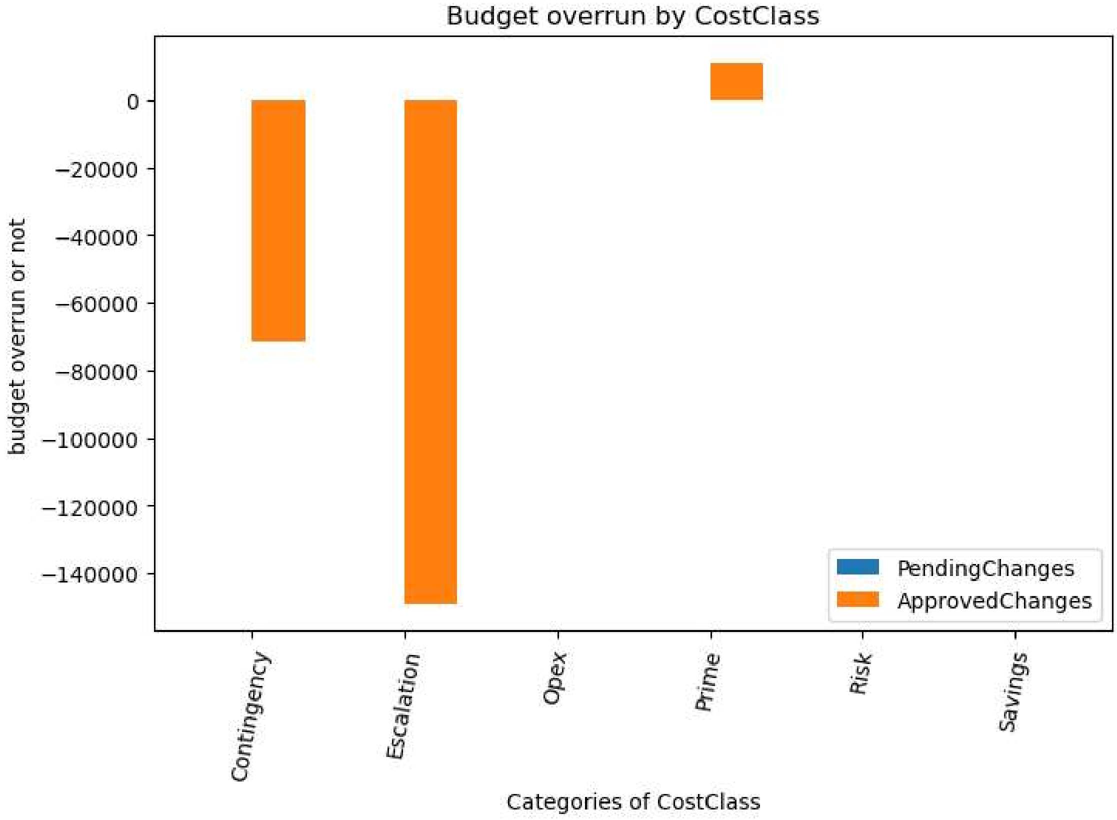
Data Understanding

Include 1 or 2 visuals that demonstrate the relationships between variables and the target variable.

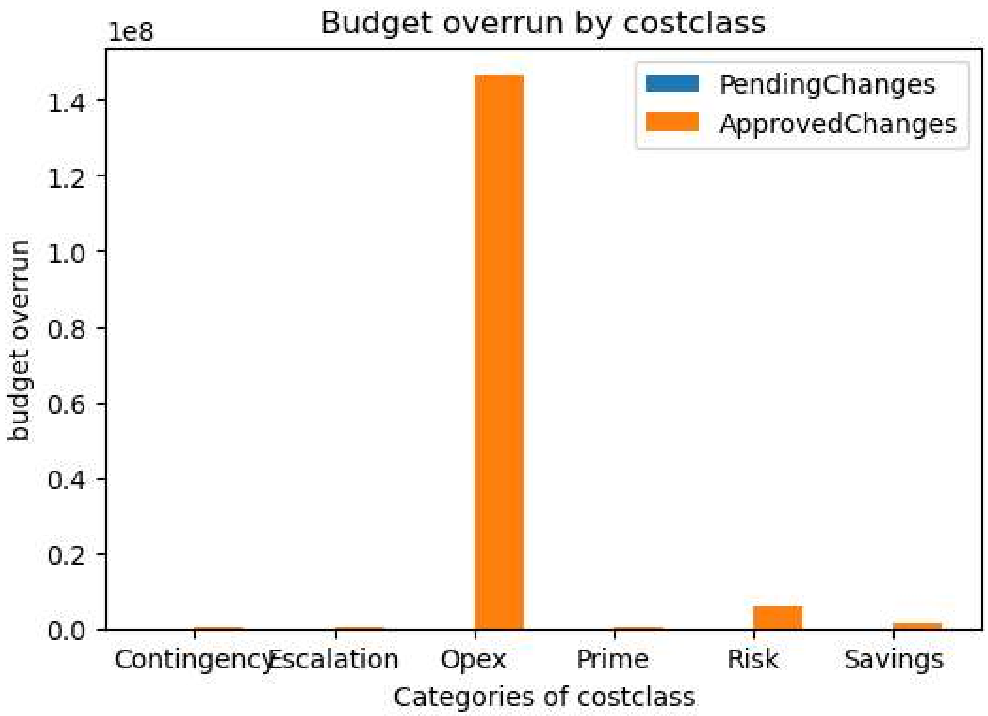
A: Here’s a visual to understand Costype and pending/approved changes either negative or positive



2. By costclass



Here’s another visual with cost class any only positive changes in pending /approved changes with the mean of pending/approved changes by each category



Did you create any other features from the dataset?

A: Yes, I created two new features from the dataset 1) Cost_change and 2) cost_overrun_value

Other Findings

Have you found any other valuable insights from the data?

A: Here are some findings from the dataset:

1) Cost overrun items in the dataset:
- No_change = 31425 (~48%)
- Negative_change = 17255 (~26%)
- Positive_change = 16833 (~25%)
2) Most of the values where there's a positive change In the original budget, are having contractor name missing
3) Approved changes in budget correlates with the overall cost overrun of the items
4) 77 package codes and 740 Items are in pending or approved changes of cost
5) 49 element codes having any type of cost changes
6) Transaction types like 'CN' and 'CO' are having some pending or approved changes in cost.
7) In cost-related features no types of charges overlap with each other

Recommendations

What do you recommend as the next steps?
A: I would suggest providing more information about what kind of projects were done which generated this dataset

- It is good to have a domain expert who can provide specific questions to address and build solutions accordingly.
- It is also recommended that datasets should be collected by each project separately with a common set of features

What information or data is missing that would be needed to develop a machine-learning solution?
A: It would be better if we could have detailed data dictionaries and background information about these datasets.