**Course - Business Intelligence Project**

**Title**- Data Driven Project Management

***Program*** *- Business Intelligence and Systems Infrastructure*

***College*** *- Algonquin College of Arts & Technology* Ottawa, Ontario

**Client** – Canada Revenue Agency



**Abstract:**

Project Management involves creating a project plan with cost estimates which may not be consistent with the actual cost for execution of said project plan due to external factors. The client’s requirement is to develop machine learning models to predict the monthly cost variance of WBS elements. Also, develop visualizations to find reasons for cost variance of said WBS elements.

**Hypothesis**:

Project costing estimates and actuals remain consistent from business opportunity problem definition (BOPD) to execution both at the project level and with the lifecycle phase.

**Background:**

In project management, the project cost estimates often vary significantly from the actual costs incurred, leading to issues in budget planning and financial management. These discrepancies can arise due to various factors such as inaccurate initial estimates, unforeseen changes in project scope, market fluctuations affecting material and labor costs, and inefficiencies in project execution. Such cost variances can result in budget overruns, affecting the overall profitability and success of the project.

If a solution is not found to address these cost variances, it can lead to several negative consequences, including:

* Budget Overruns: Persistent cost variances can cause projects to exceed their allocated budgets, leading to financial strain on the organization and potential funding shortfalls for other projects or operations.
* Project Delays: Cost overruns often result in delays as additional time and resources are required to manage and rectify budget issues, further increasing the overall project cost.
* Client Dissatisfaction: Clients expect projects to be delivered within agreed-upon budgets and timelines. Significant cost variances can damage client relationships and harm the organization's reputation.
* Resource Misallocation: Inaccurate cost estimates can lead to poor resource allocation, with either insufficient resources allocated to critical project areas or overallocation resulting in inefficiencies and wasted expenditures.
* Financial Instability: Consistent failure to manage project costs effectively can lead to financial instability for the organization, impacting its ability to secure future projects and maintain operational stability.

To mitigate these issues, it is crucial to develop predictive models that can accurately forecast monthly cost variances for Work Breakdown Structure (WBS) elements. By leveraging machine learning techniques, organizations can gain insights into the factors driving cost variances and implement strategies to control and optimize project costs more effectively. Additionally, visualizations can help in identifying patterns and trends in cost variances, enabling better decision-making and proactive management of project budgets.

**Data Source:**

An open-source dataset has been provided by the client, published on the site National Center for Biotechnology Information by Brett Thiele, Michael Ryan, and Alireza Abbasia. The dataset name is : Developing a dataset of real projects for portfolio, program and project control management research

**Short Description of Projects:**

P01 Regional Airport Carpark. This project involved the construction of two new carparks and new vehicle access to the front of the terminal at a mid-size regional airport.

P02 Regional Landfill Cell. This project included the construction of a 2 Ha lined landfill cell and leachate collection system in regional Queensland.

P03 – Regional Bypass Road. This project included the widening and strengthening of approximately three kilometers of arterial road to allow bypass of a regional mining town for heavy vehicles.

P04 – Regional Development Road. This project included the pavement rehabilitation and repairs at multiple sites along a 27 km section of a regional arterial road servicing the local mining industry.

P05 – Rehabilitation Two Roads. This project included the rehabilitation of two roads following the 2011 Queensland floods. Works included the installation of large cross and longitudinal drainage, removal and replacement of saturated subgrades, strengthening with an unbound gravel overlay, bitumen seal and asphalt.

P06 – Urban Street Rehab. This project included the reconstruction of multiple urban streets following the 2011 Queensland floods. Works included improvement of cross and longitudinal drainage, removal and replacement of saturated subgrades, replacement of all kerbing, new gravels, bitumen, and asphalt seal.

P017 - Marina Subdivision – Detailed description of projects has not been provided

P018 – Rural Road Repairs – Detailed description of projects has not been provided

**Data Characteristics of relevant parts of data:**

Each project data has similar structure and entire dataset has not been used for the analysis. The Excel worksheets relevant to analysis are Portfolio Worksheet and Actual Cost sheet from each project excel file starting with “P0” prefix which has been used for Machine learning and data visualization.

**Portfolio WBS worksheet columns:**

A standard term used in this data is quantity which is units of work.

* Portfolio WBS: The WBS element name for different tasks in project.
* Description: The description of the WBS element.
* Quantity: The total Planned Quantity of work to be completed in WBS element in a project.
* UOM: The unit of measurement for Quantity.
* BAC: Budget at Completion assigned to the WBS Element in a project.
* BAC Rate: The rate at which Budget for WBS element should increase depends on Quantity.
* Final Qty: The total Actual Quantity in units of work completed in WBS element.
* Final Cost: The total cost of WBS elements throughout the project.
* Final Rate: The rate at which Budget for WBS element really increased depending on total Actual Quantity.
* PQ: Planned Quantity of work to be completed for a WBS element each month.
* PV: Planned Value as in budget for a WBS element each month.
* AQ: Actual Quantity of work completed for a WBS element each month.
* AC: Actual Cost for a WBS element each month.
* EV: The value earned in terms of expenditure based on Actual Quantity of work completed.

The PQ, PV, AQ, AC, and EV columns are accumulated monthly and have been reformatted to have non-accumulated values.

**Actual Cost Worksheet columns:**

* Date: Date of expenditure on resource
* Description: Invoice number with description of resource if available
* Resource: Resource name
* Unit: Unit of measurement for quantity of resource
* Quantity: Quantity of resource
* Rate: Individual cost for a unit of resource
* Total: Total cost of resource spends on date.
* Portfolio WBS: The WBS name in which this resource was used.

**Executive Summary:**

* The data from 8 construction projects has been reformatted and integrated into a final dataset.
* The original dataset has each record with overall cost of a WBS element which are accumulated monthly with each project having varying duration.
* However, to fulfill client’s requirement of predicting monthly cost of WBS elements the datasets have been reformatted for monthly cost of WBS elements and difference from previous months has been calculated to get actual values instead of accumulating values.
* The integrated dataset had records where the main metrics for the machine learning model namely: Planned Quantity, Planned Value, and the target of the analysis actual cost was missing, and these records were removed or not fed to machine learning model. But they were included in data visualization.
  + The records where planned quantity, and planned value are 0 but there is actual cost associated with it are months where no work is planned but it maybe months where work is done without planning it originally.
  + The record where the actual cost is 0 is months where no work is done, therefore there is no cost associated with these months.
* After removing records which won’t help the model, during train test split about ten WBS elements had work done only once across 8 projects. But as the train test split with stratification needs at least two instances of each class. These records had to be dropped.
* The final dataset had 690 records which was given to the model.
* Machine learning models which work with distances such as Linear regression, Support Vector Machine one-hot encoding was used so the model does not make assumptions. And for models which don’t work with distances Label Encoding was used to make it less computationally heavy.
* To make the results of the model more interpretable, a 15% threshold was used to decide if the model’s prediction is accurate or wrong. And we found that the model was making 24.6% of accurate predictions for 15% threshold.
* The client requested for including additional data to be included in the data visualization which was not used for machine learning.
* This additional data had more details cost broken down by resource in each WBS elements. This was not used for machine learning because the granular cost of WBS elements would not be known before execution of WBS elements.
* Finally, A data model was developed in Power BI after integrating the results of Machine learning models and the additional data requested by client.
* The results and recommendations of the analysis are further described in the next sections of the report.

**Features used for Machine Learning Models:**

* Independent Feature: Quantity, BAC, BAC Rate, Planned Quantity, Planned Value, Portfolio WBS, UOM, Month
* Dependent Feature: Actual Cost

**Correlation of Features:**

A diagram of a number of values

Description automatically generated with medium confidence

BAC and Planned Value have strong positive correlation with target variable.

**Machine Learning Models used and the accuracy metrics:**

A white background with black text

Description automatically generated

* Linear Regression

A black and white text on a white background

Description automatically generated

* Support Vector Machine

A white background with black text

Description automatically generated

* Decision Tree

A white text with black text

Description automatically generated with medium confidence

* Random Forest Regression

The client’s requirement is to have predictions within 15% of actual cost based on this the accuracy of the models is:

A white paper with black text and numbers

Description automatically generated

For 15% tolerance the best performing model is SVM with 24.64% accuracy.

**Analysis & Explanations:**

A data model was developed in Power BI to explore the results and communicate the insights from analysis.

A screenshot of a computer

Description automatically generated

Multiple slicers have been developed to explore the data

A screenshot of a computer screen

Description automatically generated

The first visual aims to show the difference between the actual cost and predicted cost using the four models. Random forest regression has the closest overall prediction, but the accuracy of Support Vector Machine is highest for individual months.

A graph of different colored lines

Description automatically generated with medium confidence

The second visual aims to compare monthly predictions of four models with actual cost. Since the data has many WBS elements which are categorical. It could be a model that predicts accurately for some WBS elements even if its overall accuracy is low. So, the best performing model can be found for a specific WBS element.

A screenshot of a computer screen

Description automatically generated

The third visual is a descriptive analysis to find the top 20 resources on which have incurred highest actual cost.

A screenshot of a computer

Description automatically generated

The fourth visual aims to show the most expensive resources.

A screen shot of a graph

Description automatically generated

The fifth visual aims to show the top 20 WBS elements which have the highest cost variance.

A screenshot of a graph

Description automatically generated

This visual is important, it shows a big decision in our analysis. The left visual compares Actual cost and Planned value for original dataset. The right visual compares the same but only for the data which has been used for Machine Learning. This is because as we previously mentioned the data had months where there is no budget allotted for a month so, Planned Value and Planned Quantity is zero, but Actual cost is associated with the month. It’s clear when no work is planned but cost is associated with the month such data should not be given to the model.

A screenshot of a graph

Description automatically generated

This visual shows why the model is underpredicting actual cost, it’s because the model was given Planned Quantity as input, which is planned work, but Actual Quantity is not given to model as at the time of prediction it would not be known. The cost depends on work done as a client would not pay the full amount when the planned work is not completed.

A screenshot of a computer screen

Description automatically generated

This visual shows a count of WBS items and the degree of cost variance they have when going over the budget. As can be seen a total of 158 times WBS elements have gone over the budget with huge cost variance.

A screenshot of a graph

Description automatically generated

This visual shows the Top 20 Resources which are common or mostly worked upon across projects.

A screenshot of a graph

Description automatically generated

This visual shows how many times the work targets are exceeded in left visual and in right visual it shows how many times the work target is failed to reach.

A screenshot of a graph

Description automatically generated

This visual shows effect on Actual cost when work target is exceeded. Since the model only knows Planned Quantity and Planned Value it tries to make predictions based on that its predictions fail since it doesn’t know more work is completed leading to increase in actual cost.

A screenshot of a graph

Description automatically generated

This visual shows the effect on actual cost when work target is not reached. Interestingly, the model tries to make accurate predictions by making predictions for the case when work target is not reached by predicting cost closer to actual cost.

The model takes the case of not reaching the work target as priority to increase accuracy as in the dataset majority case is not reaching the work targets.

**Conclusion & Recommendations:**

In conclusion, the model is trying to make accurate predictions for the case when work targets are not reached. An additional metric can be included which shows the company or team working on the project. This can be useful as there could be a pattern in a team usually exceeding or failing to reach the work target. Based off this the model can make accurate predictions. Additionally, duration can be included but this still may not be most accurate as duration is also a metric which can fluctuate.

**References:**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7770534/>

<https://figshare.com/articles/dataset/Project_Portfolio_Dataset/12998822?file=24779921>

Thiele B, Ryan M, Abbasi A. Developing a dataset of real projects for portfolio, program and project control management research. Data Brief. 2020 Dec 17;34:106659. doi: 10.1016/j.dib.2020.106659. PMID: 33385027; PMCID: PMC7770534. [Accessed July. 21, 2024]