**Capstone Project Report**

**on**

**Forecasting Project Cost Variance and Identifying Project Success Factors**

Towards partial fulfilment of course

CST2212-Business Intelligence Project

Table of Contents

[Table of Contents 2](#_Toc164425172)

[Executive Summary 3](#_Toc164425173)

[Problem Statement 4](#_Toc164425174)

[Logical Approach & Methodologies 5](#_Toc164425175)

[Analysis & Findings 10](#_Toc164425176)

[Machine Learning 14](#_Toc164425177)

[Literature Review 23](#_Toc164425178)

[Future Work & Recommendations 25](#_Toc164425179)

[Conclusion 26](#_Toc164425180)

[Lessons Learned 27](#_Toc164425181)

[References 28](#_Toc164425182)

[Appendix 1 30](#_Toc164425183)

[Appendix 2 31](#_Toc164425184)

# Executive Summary

In this Project, we have taken a significant stride towards revolutionizing project management strategies by developing a sophisticated model aimed at predicting project cost variances and success rates. The primary objective of this initiative is to refine the accuracy of cost estimates and enhance the predictive analytics capabilities surrounding project outcomes. By leveraging historical data and employing advanced machine learning techniques, the team has identified predictive models and factors that not only help forecast cost fluctuations but also identify critical success factors for projects.

The project embarked on a detailed analysis of past project data, which facilitated a deeper understanding of the elements that contribute to the success and failures of projects. Utilizing tools such as Python, Scikit-learn for machine learning model implementation, and Power BI for data visualization, the team implemented several algorithms including Linear Regression, Random Forest, and MLP Regressor. These models were optimized to predict project variables effectively, which in turn supports more reliable and financially sound project management.

Furthermore, the project aligns with the client's needs by offering actionable insights that are crucial for making informed decisions and improving overall project oversight. The insights derived from the machine learning analysis enable the client to manage projects with greater efficacy, ensuring projects are completed on time, within budget, and according to specifications.

This initiative not only contributes to the theoretical field of project management through its exploration of predictive analytics but also provides practical solutions that can be implemented in real-world scenarios. As such, the project stands as a testament to the potential of integrating machine learning and data analytics into project management, paving the way for more dynamic and successful project execution.

# Problem Statement

The project aims to improve project cost management through predictive modeling and advanced data visualization, empowering stakeholders for better decision making and resource allocation. The problems statement includes:

1. To develop predictive models to accurately forecast project cost variances.
2. Data visualization to analyze and present these variances, enabling efficiencies and decisions throughout the project lifecycle.

**Predictive Models for Project Cost Variances:**

The objective is to develop a forecasting model that can accurately predict costs that vary within the project budget. This model will use historical data, project metrics, and machine learning algorithms to predict deviations from budgeted costs. By anticipating cost variances, project managers can anticipate potential budget overruns and optimize resource utilization.

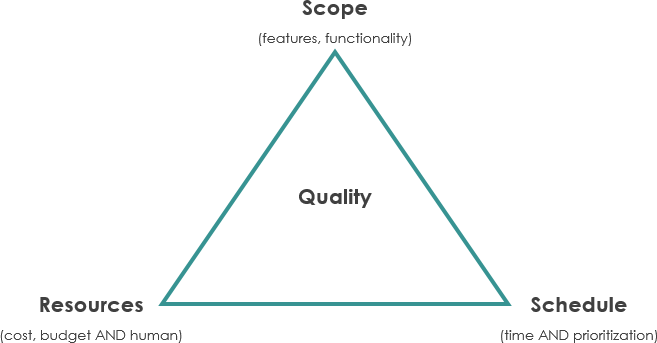
**Data Visualization Tools for Cost Variance Analysis:**

Develop flexible data visualization tools that will enable stakeholders to analyze cost variance and interpret cost variance throughout the project lifecycle more effectively. These tools provide visualizations and interactive dashboards, allowing users to analyze trends, identify patterns, and make informed decisions. By effectively introducing cost differentials, project teams can increase transparency, facilitate communication, and increase efficiency in project management.

# Logical Approach & Methodologies

The basis of the logical solution of our project was built on a critical examination and understanding of the main principles of project management, especially through the optics of the Iron Triangle - scope, time, and costs. This traditional framework, while essential, has been expanded in our approach to include a broader spectrum of factors that significantly influence project outcomes in the current environment.

**Iron Triangle:**

[](https://www.visual-paradigm.com/servlet/editor-content/project-management/what-is-iron-triangle-of-projects/sites/7/2019/09/the-iron-triangle.png)

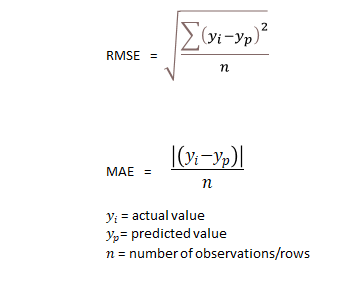
*Figure 1[1]*

It represents infrastructure project management principles that focus on scope, time, and cost. It has been expanded to include additional key dimensions: Effort Cost', 'Actuals', and 'Capacity to Deliver', providing a comprehensive view of project dynamics beyond traditional constraints[[4].](#_References)

Methodologies:

The transition from our logical approach to tangible methods and solutions is driven by the careful application of various machine learning models.

**Model Evaluation Metrics**

[](https://miro.medium.com/v2/resize:fit:710/1*5OQunI-NR-S0gAZFIit1Rw.png)

*Figure 2[2]*

In the evaluation phase of our project, we used two main metrics to evaluate the performance of our prediction model: mean squared error (MSE) and root mean squared error (RMSE) These metrics are important because they distinguish a model between predicted values and observed actual values. Using MSE, we can estimate the squared difference, which provides a granular view of model performance. RMSE, as the square root of MSE, provided a highly interpretable metric in the same units as the target variable. These metrics were important in fine-tuning our model and enabled us to repeatedly improve our forecasts to better match real-world results and ensure that our price volatility forecasts were achievable, reliable, and applicable in useful contexts [[5].](#_References)

**Machine Learning Model Approach**

1. **Linear Regression Model:**

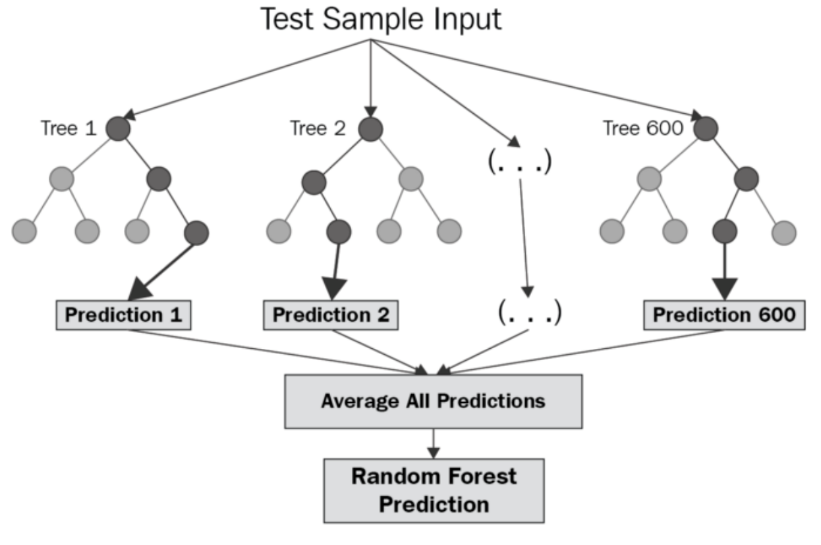
**[A graph with blue dots

Description automatically generated](https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/400px-Linear_regression.svg.png)**

*Figure 3[3]*

We Used Linear Regression, a basic model that excels at predicting continuous outcomes and can provide a starting point for performance. Its simplicity is useful for illustrating the direct effect of the forecast variables on our target.

1. **Random Forest Regressor Model:**

[](https://miro.medium.com/v2/resize:fit:1400/1*ZFuMI_HrI3jt2Wlay73IUQ.png)

*Figure 4[4]*

We used a Random Forest Regressor to capture potential complex, nonlinear relationships in the data. This cluster model with multiple decision trees is adept at handling overfitting and provides a more nuanced understanding of dynamic interactions.

1. **Multilayer Perceptron (MLP) Regressor Model:**

**[A diagram of a network

Description automatically generated](https://miro.medium.com/v2/resize:fit:1400/1*-IPQlOd46dlsutIbUq1Zcw.png)**

*Figure 5[5]*

For even more complex data models, a Multilayer Perceptron (MLP) Regressor was introduced. As a form of neural network, MLP has potential for capturing deep nonlinear relationships, making it an excellent choice for modeling dynamic changes where it is not easy to formulate the relationship to a traditional linear one in various examples.

**Model Optimization and Hyperparameter Tuning**

**[A diagram of a model

Description automatically generated](https://media.licdn.com/dms/image/D4D12AQHFJPGKuPjmDg/article-cover_image-shrink_600_2000/0/1685739097091?e=2147483647&v=beta&t=qXmPm8UIteILTcuCoLzOKYeTGJTo17uoHdt2z2e4dZ8)**

*Figure 6[6]*

Refining our machine learning model in project prediction is an important step towards achieving accuracy. Starting our analysis with basic models such as linear regression, we resorted to more complex random forest and MLP regressors. Through grid search and cross-validation, we carefully optimized the model parameters, significantly increasing their prediction accuracy. This not only sharpened our predictions but also deepened our understanding of the complex structure of the dataset, thereby improving our predictions of cost variance forecasts.

**Data Analysis Tools (Power BI):**

**A screenshot of a project management dashboard

Description automatically generated**

*Figure 7*

Our work has been supported by Power BI, which acts as a visualization tool that brings our complex data to life. It provides a platform for transforming our large number of data points into interactive, visually appealing, facilitating easy interpretation of project financial metrics. With its intuitive design and rich feature set, Power BI enables us to build a narrative of project costs, gaps, and other important metrics, providing stakeholders with a clearer understanding there and immediately on the financial health of the business. Through custom dashboards, we can view real-time data, analyze specific trends with just a few clicks, and make appropriate decisions quickly and efficiently.

Some of the analysis is also done in Python, Scikit-learn, a library known for its machine learning tools. This integration enabled us to create and analyze patterns that shaped our data- driven insights.

**Jira for Project management**

Jira, our chosen project management platform, provided a structured and transparent framework for tracking our project's progress. It ensured that every team member was aligned with the project's goals and timelines, fostering a collaborative and efficient working environment.

In summary, our project's logical and physical solutions represent a blend of theoretical understanding and practical application. By integrating advanced machine learning models, data visualization, and project management tools, we have developed a comprehensive framework that not only forecasts project cost variances but also provides strategic insights into the factors that drive project success.

**Correlation Matrix:**

A chart with different colored squares

Description automatically generated

*Figure 8*

Correlation heat map provides a straightforward visual guide to understanding how different elements of the project are related. For example, our project concept of the positive correlation between 'effort cost' and 'authenticity' shows a direct relationship where increased effort is reflected by an increase in the actual cost. Furthermore, the way in which ‘priority’ corresponds to this cost reflects the impact of project urgency on budgeting. At the other end of the spectrum, the inverse relationship between ‘actual’ and ‘profit margins’ suggests that the increase in debt may match the planned budget. This heat chart is a useful tool for identifying the aspects of the project that most affect the financial results.

The backbone of our logical solution is advanced predictive modeling algorithms. This process was not only a mathematical exercise but also a way of trying to extract meaningful patterns and insights from the historical project data. We used sophisticated machine learning techniques to aim to predict price differentials and identify high accuracy success indicators This includes a rigorous process of hypothesis testing, model selection and iterative adjustment to ensure that our forecasts are reliable and relevant.

# Analysis & Findings

Diving deep into comprehensive examination of the data collected and the insights derived from the exploratory data analysis and the correlation matrix. The primary focus was to determine relationships and patterns that could influence project cost variances (CV) and overall project success. These findings will influence the major factors that can be used as input variables for a CV machine learning model.

To understand the data better, we decided to build a dashboard that can showcase and highlight the information in a clean and visually appealing manner. We have used Power BI as our main tool for dashboard creation and doing analysis. This helped us lay out the available data in a concise manner.

**Dashboard**:

**A screenshot of a project management dashboard

Description automatically generated**

*Figure 9 (Projects View)*

The dashboard is divided into 2 different tabs to display and convey a specific story based on user requirements. The figure above shows the “Project View” which dives deeper into and filtered by historical projects. Some of the charts give information like Priority, Work Classification, IT Directorate, Cost Variance and Budget classifications. This information can enable a project manager to investigate each project in depth and determine problem areas and bring improvements.

Custom measure “Budget” was added to the data to identify which projects, either ongoing or finished, are Over/under Budget. To achieve this, we have used CPI = Efforts/Actuals and then categorized based on if CPI > 1 = Under budget, CPI < 1 = Over Budget.

A screenshot of a project management dashboard

Description automatically generated

*Figure 10 (Organization View)*

Organization View of the dashboard gives insight into “The people” working on the said projects, since employees and managers are a factor in contributing to a project’s success or failure. This view is divided into charts that give insight into Work Classification, Projects, and Cost Variance of a project by IT Directorate, IT Division, and IT section.

**Findings:**

Based on the dashboards and data provided, we have identified some interesting findings and factors that directly affect the cost variance of a project, and later can be used as input variables of a machine learning prediction model.

1. **Priority:**

It becomes evident that Priority has the most significant impact on Cost Variance (CV). Projects categorized as ‘Low’ and ‘Medium’ Priority exhibit a tendency to exceed their budget allocations. Conversely, ‘Critical’ Priority projects demonstrate a tendency to remain within or slightly surpass budgetary constraints.

A screenshot of a graph

Description automatically generated

*Figure 11*

1. **Work Classification:**

The analysis indicates that Cost Variance (CV) does not strongly correlate with Work Classification. However, it is noteworthy that projects categorized as 'Enhancement' exhibit the highest CV among all project types. It also goes together with “Priority” as well since highest variance is seen in enhancement applications with medium priority and development applications with low priority.

1. **IT Directorate:**

**A screenshot of a graph

Description automatically generated**

*Figure 12*

The organization view analysis dictates that the Team, Section/Division working on the project also plays a role in affecting cost variance. Based on the data, we derived a trend of certain IT Team always delivering projects over budget with very high-cost variances. This can be because of many factors, just to name a few, lack of experience and skills, poor project planning, etc. For example, IT Dir -1 has historically delivered always delivered projects over budget, meanwhile IT Dir 3&6 has all their projects under budgets.

A screenshot of a computer screen

Description automatically generated

*Figure 13*

Based on the findings of crucial factors affecting CV, same factors can also contribute to a successful project well under budget with a proper project planning, providing proper priority and schedules based on needs, evaluating the team on their skills and experience that would be working on the projects.

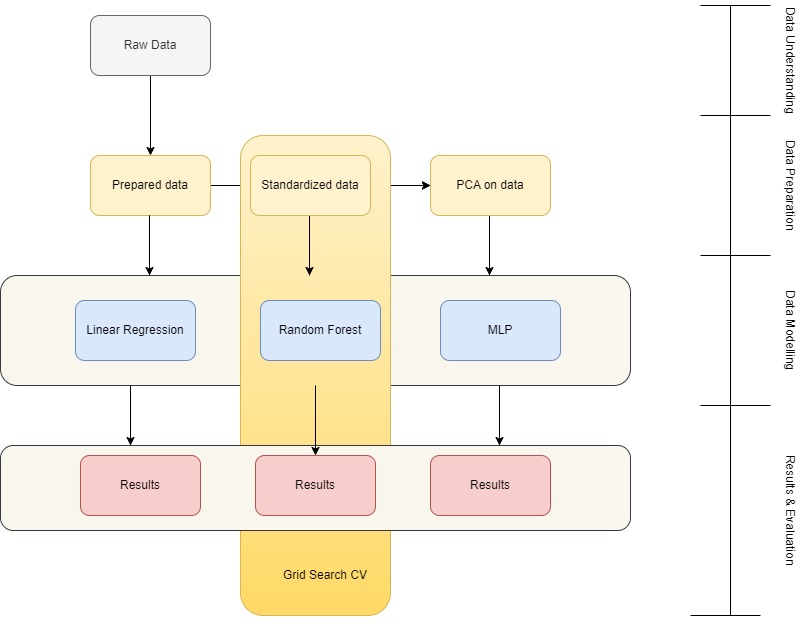
# Machine Learning

Machine learning is a branch of artificial intelligence that focuses on creating algorithms capable of learning from data to make predictions or decisions without being explicitly programmed. It involves training models using labeled data, where both input features and corresponding target labels are known, to learn patterns and relationships. There are several types of machines learning algorithms, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves learning from labeled data to make predictions or classify inputs into categories. Unsupervised learning deals with discovering patterns and structures in unlabeled data. Reinforcement learning is about training agents to make sequential decisions by interacting with an environment to maximize rewards. Machine learning finds applications in various domains, including healthcare, finance, natural language processing, computer vision, and more, where there is a need to automate tasks, make predictions, or gain insights from data [12].

This section will focus on the approach that followed during the machine learning processed in the problem statement. This part focuses on developing a forecasting model that can accurately predict costs that vary within the project budget. This model will use historical data, project metrics, and machine learning algorithms to predict deviations from budgeted costs. By anticipating cost variances, project managers can anticipate potential budget overruns and optimize resource utilization.

Machine Learning section is divided in to main 4 sub parts as follows:

1. Data Understanding
2. Data preparation
3. Data Modelling & Model tunning
4. Results & Evaluation



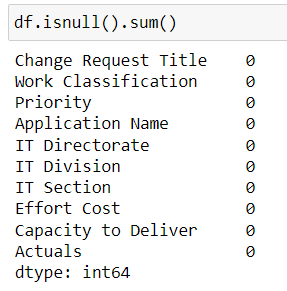
*Figure 14*

1. **Data Understanding:**

Data understanding involves exploring and understanding each feature and its datatype, one can understand the entire project sharply by validating the data quality and make further improvement according to their use and goal of the analysis. Dataset contains 64 instances.

|  |  |
| --- | --- |
| **Attribute** | **Attribute type** |
| Order Number | int64 |
| Change Request Title | str |
| Work Classification | str |
| Priority | str |
| Application Name | str |
| IT Directorate | str |
| IT Division | str |
| IT Section | str |
| Manager ID | int |
| Effort Cost | float |
| Capacity to Deliver | str |
| Actuals | float |

Verified that there are no null values in data set.



*Figure 15*

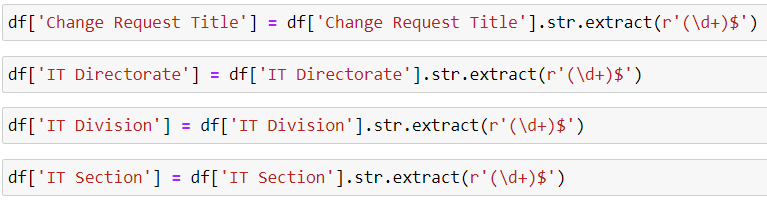
1. **Data preparation:**

Data preparation is the crucial step in the data analysis process, involving transforming raw data into a format suitable for analysis. This includes tasks such as cleaning the data to remove errors and inconsistencies, handling missing values through imputation or deletion, and restructuring the data to meet the requirements of the analysis. Additionally, data preparation often involves standardizing data formats, scaling numerical features, encoding categorical variables, and feature selection or extraction to enhance the quality and relevance of the data for subsequent analysis tasks. Efficient and thorough data preparation lays the foundation for accurate and meaningful insights to be derived from the data.

Following attributes were **removed** from dataset as it was not giving any benefit for machine learning process:

* Order Number
* Manager ID

Following attributes were converted into their suffix value format entirely using given technique:



*Figure 16*

|  |  |
| --- | --- |
| **Before** | **After using above technique** |
|  |  |

In the following process, new feature was created using existing features:



|  |
| --- |
| Output |
|  |

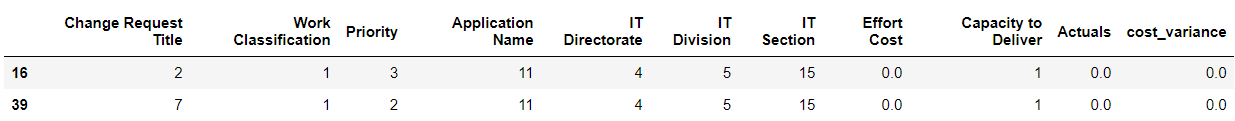
Followed by that, label encoding was used for converting categorical values to numeric values, for example, converting A, B, C, D to 0,1,2,3.

|  |  |
| --- | --- |
| Before | After using Label Encoding |
|  |  |

* **Outlier Detection and Removal:**

Outlier detection is a data analysis technique aimed at identifying data points that significantly deviate from the rest of the dataset. These outliers may represent anomalies, errors, or rare events within the data. By detecting outliers, analysts can gain insights into potential data quality issues, identify abnormal behavior, and make informed decisions about data preprocessing, anomaly detection, or further investigation. Various statistical methods, machine learning algorithms, and visualization techniques are employed to detect and analyze outliers in different domains, including finance, healthcare, and cybersecurity.

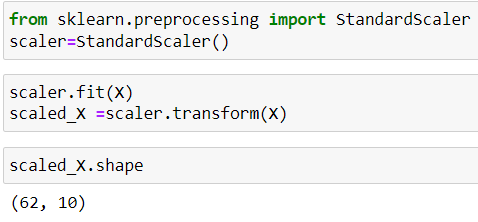
Here in our case, capacity to deliver = 1 & cost variance=0, these instances considered as an outliers and was removed from dataset.



* **Standardize Data:**

Standardization is a process used to transform data so that it conforms to a common scale or format. This ensures that different datasets or variables can be compared or combined more easily. By standardizing data, we remove variations in units or scales, making it simpler to analyze and interpret results across different contexts or measurements.

In this case, we used standard scaler for scaling:



*Figure 17*

* **PCA:**

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction in data analysis. It aims to transform a dataset into a new coordinate system where the variables are uncorrelated and ordered by their variance. By identifying and retaining the most important information while discarding less relevant details, PCA simplifies complex datasets, making them easier to visualize and analyze. This reduction in dimensionality can help uncover underlying patterns, relationships, and structures within the data.

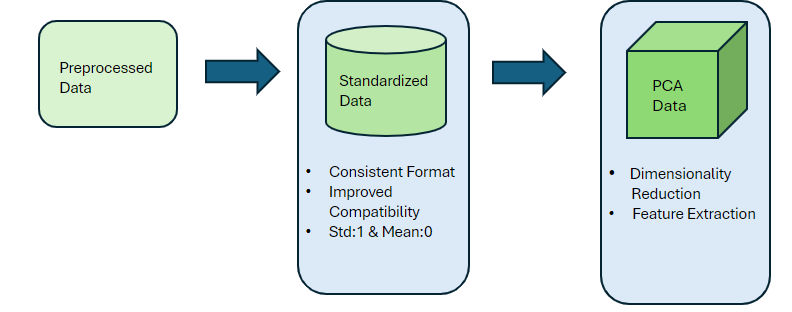
|  |  |
| --- | --- |
| This graph (Elbow Graph) is used to determine Principal component as per out dataset, which will further modified at the same place in code to work with the same component in modelling. |  |

* **Train-test Split:**

Train-test split is a common technique used in machine learning to evaluate the performance of a model. It involves dividing the dataset into two subsets: one for training the model and another for testing its performance. Typically, the training set is used to train the model on known data, while the testing set is used to evaluate how well the model generalizes to unseen data. This practice helps assess the model's performance, such as its accuracy or error rate, before deploying it in real-world scenarios, thus providing insights into its reliability and effectiveness.

|  |
| --- |
| from sklearn.model\_selection import train\_test\_split    # Split the original dataset into train and test sets  X\_train\_scaled, X\_test\_scaled, y\_train\_scaled, y\_test\_scaled = train\_test\_split(scaled\_X, y, test\_size=0.3, random\_state=42)  X\_train\_pca, X\_test\_pca, y\_train\_pca, y\_test\_pca = train\_test\_split(pca\_X, y, test\_size=0.3, random\_state=42) |

**Flow of the data preparation:**



*Figure 18*

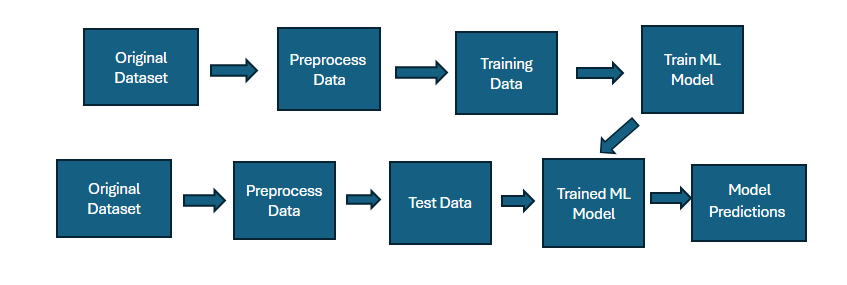
As per the above image, 3 types of datasets were prepared to apply on models.

1. Preprocessed data
2. Standardized data
3. PCA data
4. **Data Modelling & Model tunning:**

Modeling in the context of machine learning involves the creation and refinement of algorithms or mathematical representations that can learn patterns and relationships from data. This process typically begins with selecting an appropriate model architecture or algorithm based on the nature of the problem and the available data. Next, the model is trained using a subset of the data, during which it learns to make predictions or classifications based on the input features [12].

Once trained, the model is evaluated using a separate dataset to assess its performance and generalization ability. This evaluation helps identify any issues such as overfitting (where the model performs well on the training data but poorly on new data) or underfitting (where the model is too simplistic to capture the underlying patterns in the data). The model may then be fine-tuned or adjusted based on the evaluation results to improve its performance. After the model has been trained and validated, it can be deployed to make predictions or classifications on new, unseen data. This final step completes the modeling process, allowing the model to be used in real-world applications to automate tasks, make decisions, or generate insights.

Flow of the modelling process can be determined by following image:

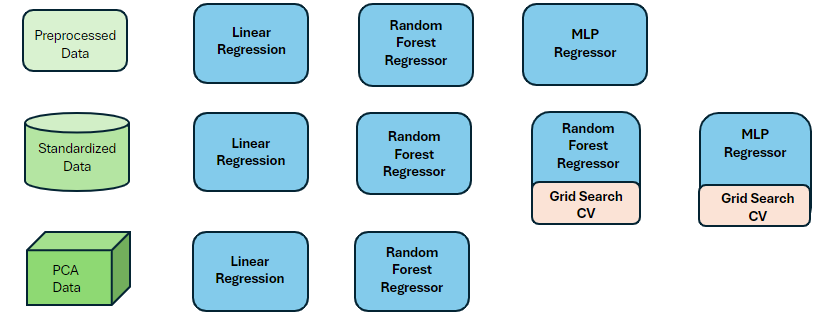


*Figure 19*

As mentioned earlier, 3 modeling techniques were used on all 3 prepared datasets. As we are predicting continuous features using all numeric values (which have been prepared for a purpose to fit in regressor models), all regression models were used for modeling [13].

1. Linear Regression
2. Random Forest Regressor
3. Multilayer Perceptron Regressor

The visual representation of above statement can be presented as mentioned below:



*Figure 20*

* **Hyperparameter tunning using grid search CV:**

Hyperparameter tuning is a critical phase in machine learning model development where the optimal configuration of hyperparameters is determined to enhance model performance. Hyperparameters are parameters that influence the learning process of a model but are not directly learned from the data. This tuning process aims to strike a balance between model complexity and generalization by fine-tuning hyperparameters such as learning rate, regularization strength, or tree depth. By systematically exploring different hyperparameter settings, practitioners can uncover configurations that lead to improved model performance, higher accuracy, and better generalization to unseen data.

Grid search cross-validation (GridSearchCV) is a widely used technique for hyperparameter tuning that systematically explores a predefined grid of hyperparameter values. It automates the process of searching through the hyperparameter space by evaluating each combination using cross-validation. GridSearchCV exhaustively searches through all possible combinations of hyperparameters, providing a comprehensive approach to finding the optimal configuration. By evaluating each combination's performance using cross-validation, GridSearchCV helps mitigate the risk of overfitting and ensures that the selected hyperparameters generalize well to unseen data. This systematic approach to hyperparameter tuning is invaluable in optimizing model performance and enhancing the overall effectiveness of machine learning models.

In this project, GridSearchCV was used on standardized data as it was proven by results that standardized data gives the best results out of any other prepared datasets.

1. **Results & Evaluation:**

Results and evaluation are crucial components of the machine learning pipeline, providing insights into a model's performance and its ability to generalize to unseen data. After training a model using the selected hyperparameters, it is essential to evaluate its performance on a separate test dataset. This evaluation allows practitioners to assess the model's accuracy, precision, recall, F1 score, or other relevant metrics, depending on the specific task. By comparing the model's predictions against the ground truth labels in the test set, practitioners can determine how well the model performs in real-world scenarios and identify any potential issues such as overfitting or underfitting.

* **Evaluation Matrix:**

1. Evaluation metrics such as **Mean Squared Error (MSE)** are essential tools for quantifying the performance of regression models. MSE measures the average squared difference between the actual and predicted values, providing a measure of the model's accuracy. A lower MSE indicates better model performance, as it signifies that the model's predictions are closer to the actual values. MSE is particularly useful when dealing with continuous target variables, allowing practitioners to assess the overall goodness-of-fit of the regression model. By comparing the MSE of different models or tuning hyperparameters, practitioners can identify the most effective model configuration for their specific regression task [12].
2. **Root Mean Squared Error (RMSE)** is a related evaluation metric that measures the square root of the MSE. RMSE provides a more interpretable measure of error in the same units as the target variable, making it easier to understand and communicate model performance. Like MSE, a lower RMSE indicates better model accuracy, as it signifies that the model's predictions have smaller errors on average. RMSE is widely used in regression tasks, offering a comprehensive assessment of model performance that accounts for both bias and variance. By considering both MSE and RMSE, practitioners can gain a holistic understanding of a regression model's accuracy and effectively compare different models or parameter configurations [12].

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Applied** | **Data Format** | **Results (MSE)** | **Results (RMSE)** |
| Linear Regressor | Prepared data | 4.4572e-20 | **x** |
| Linear Regressor | Standardize data | 3.1999e-21 | **x** |
| Linear Regressor | PCA data | 51534722817.22 | **x** |
| Random Forest Regressor | Prepared data | **x** | 239592.653 |
| Random Forest Regressor | Standardize data | **x** | 191310.620 |
| Random Forest Regressor | PCA data | **x** | **x** |
| MLP Regressor | Prepared data | 89.875 | **x** |
| MLP Regressor | Standardize data | **x** | **x** |
| MLP Regressor | PCA data | **x** | **x** |
| Random Forest Regressor with Model Tunning | Prepared data | **x** | 38802.595 |
| Random Forest Regressor with Model Tunning | Standardize data | **x** | 38802.595 |
| MLP Regressor with Model Tunning | Standardize data | 135.742 | **x** |

Out of all Models - Data combinations, Linear regression with standardized data performed the best with MSE of 3.1999 e-21 which is closed 0. As per the analysis and the modelling, this kind of preparation should be done on similar datasets on linear regression model to predict highly accurate cost variance.

# Literature Review

For the purpose of identifying other factors that help forecast project cost better, intuitively as well as for the models, we reviewed some research papers. The goal was to come up with some factors/variables/features, that can be incorporated in the dataset which can help the model better learn different patterns and improve the performance of the model. Based on the review of the research papers, we identified three such domains and related features which can be incorporated in the dataset. These are discussed further.

**1.** **Different models for various phases of project**

The research paper authored by Narbaev T. et. al [1], discuss their approach to improve the cost estimates for various projects in the construction domain with the help of machine learning. They focus on cost features such as Earned Value (EV), Actual Value (AC) and Budget at Completion (BAC) to predict Estimate at Completion (EAC).

The novel approach discussed in this research paper is the introduction of a “tracking period” feature. This feature bins the cost values based on the project completion percentage value. The 3 bins created based on the project completion percentage were: 1-30% Initial stage, 31-70% as middle stage and 71-95% as late stage. Narbaev T. et. al suggest this helps in creating separate machine learning models for various project completion stages[1]. This approach may have a great advantage as it has the potential to give accurate cost estimates to project managers at various stages of the project.

Lastly, based on the performance results of various machine learning models, models XG Boost and SVM performed extremely well compared to other regression models like linear regression. The reason for this better performance is given to robustness of the model to handle input features of various datatypes and control of hyper-parameters to improve the accuracy of the models. Even though this project is an IT project, some similar concept of project completion tracking variable can be introduced create different models for various phases of project and predict accurate cost estimations which will eventually help to make better decisions.

2. **Introduction of Risks related factors**

The second domain is related to the introduction of risk related factors. The authors Saeidlou S. and Ghadiminia N. discuss application of Deep neural networks for cost estimation construction projects [2]. The authors introduced risk related features in their dataset which has some impact on the prediction success.

Taking that notion of risk related features, some variables that can be incorporated in this dataset are discussed below. First is qualitative risk score (probability\*severity). This can help the model as well as the project manager get an understanding of a factor which impacts on the project costs. The second risk feature of great importance which impacts on the project costs, time, and scope all together is the number of change requests made/approved. As it is known that the higher the number of change requests in the project, the higher the project is at the risk of failure. Lastly, the number of risks triggered during the project phase can be introduced in the model as well. This feature can help the model understand the relationship between number of risks triggered during the project and the cost estimation based on that.

3. **Features related to communication and stakeholder engagement**

Another type of feature that can be introduced in this dataset is through various surveys related to experience of project team members and other stakeholders. This idea was introduced by the authors Bang S et al. [3]. Based on this above notion, some introductory variables that can be incorporated into the dataset are discussed further. Firstly, a variable measuring adequateness of processes required in the project/sub-project can be incorporated. As an introductory variable, a binary feature (Yes/No) stating the compliance be introduced. The second feature that can be introduced is related to training. As an introductory variable, a binary feature (Yes/No) stating whether project team members including the managers were provided with adequate training or not. Additionally, a feature related to frequency of communication of objectives and goals to team members can be introduced to let the model know the impact of communication activities on project cost estimate. Lastly, any feature that measures buy-in level of stakeholders towards objectives/goals can be incorporated into the dataset as well.

Thus, the introduction of features related to various aspects of project management has the potential to improve the forecast of project cost estimates. Various attempts should be made to collect as much information related to the project as possible.

# Future Work & Recommendations

Based on all the discussion and analysis provided previously, the following areas will have the most impact on improving the project cost estimation and overall project success.

* Root Cause Analysis for over-budget reasons should be performed for projects with Priority “Medium” and “Low”, followed by solution preparation of causes identified based on Pareto principle.
* Processes for estimation of “effort cost” should be revisited for improvement purpose.
* For accurate predictions of cost variance, additional factors related to risks, training, communication, team engagement should be included.
* Additional Machine Learnings models like XGBoost, Support Vector Machines should be tried and compared with previous results.

# Conclusion

Based on all the analysis and discussion, following things can be concluded for the project success prediction:

* Project success prediction based on cost alone has limited value. Better predictions can be made with the addition of extra features.
* Simple prediction models like Linear Regression perform better than complex prediction models like Neural Networks when cost factor is involved.
* Priority and Organizational structure has some influence over the projects being over budget.

# Lessons Learned

Following lessons were learned during the execution of the project as well as the project related activities:

* Collaboration is extremely important for successful execution of the project.

* 15-30 minutes meetings should be carried out to discuss problems faced and work on solutions for problems faced, collectively, every day to resolve issues quickly.
* Various tips and tricks related to Jira and GitHub should be learnt prior to start of the project to make better use of the tools.
* Clear, concise and easy to understand task preparation for JIRA should be carried out.
* If using GitHub, the files and folder names should be consistent along with code structures, following best practices, to minimize the conflict.
* To facilitate faster understanding of data, a data dictionary should be created with the help of stakeholders.
* Before preparation of content for presentation, audience's level of understanding about the project should be gathered to know what granularity of detail is needed or what concepts should be covered.

# References

1. Timur Narbaev, Öncü Hazir, Balzhan Khamitova & Sayazhan Talgat (2023) A machine learning study to improve the reliability of project cost estimates, International Journal of Production Research, DOI: [10.1080/00207543.2023.2262051](https://doi.org/10.1080/00207543.2023.2262051)
2. Salman Saeidlou & Nikdokht Ghadiminia (2024) A construction cost estimation framework using DNN and validation unit, Building Research & Information, 52:1-2, 38-48, DOI: 10.1080/09613218.2023.2196388
3. Bang S, Aarvold M O, Hartvig W J, Olsson N O E, Rauzy A (2022). Application of machine learning to limited datasets: prediction of project success, ITcon Vol. 27, pg. 732-755, <https://doi.org/10.36680/j.itcon.2022.036>
4. Project Management Institute. (n.d.). The Triple Constraint. Retrieved from <https://www.pmi.org/learning/library/triple-constraint-10260>
5. Acharya, S. (2021). What Are RMSE and MAE? Towards Data Science. Retrieved from [https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383](https://towardsdatascience.com/what-are-rmse-and-mae-e405ce2303831)
6. Project Management Institute. (2017). A guide to the project management body of knowledge (PMBOK® guide) (6th ed.). Project Management Institute.
7. Nair, V. (2021). APA referencing guide for the built environment.
8. Routledge.Kerzner, H. (2017). Project management: A systems approach to planning, scheduling, and controlling (12th ed.). Wiley.

**Figures:**

1. “The-Iron-Triangle.png (657×343).” <https://www.visual-paradigm.com/servlet/editor-content/project-management/what-is-iron-triangle-of-projects/sites/7/2019/09/the-iron-triangle.png>
2. RMSE and MAE.png[https://miro.medium.com/v2/resize:fit:710/1\*5OQunI-NR-S0gAZFIit1Rw.png](https://miro.medium.com/v2/resize:fit:710/1*5OQunI-NR-S0gAZFIit1Rw.png)
3. Wikimedia Commons. (n.d.). Linear regression <https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/400px-Linear_regression.svg.png>
4. Miro. (n.d.). Random forest regression [Image]. [https://miro.medium.com/v2/resize:fit:1400/1\*ZFuMI\_HrI3jt2Wlay73IUQ.png](https://miro.medium.com/v2/resize:fit:1400/1*ZFuMI_HrI3jt2Wlay73IUQ.png)
5. Miro. (n.d.). Multilayer perceptron (MLP) regressor [https://miro.medium.com/v2/resize:fit:1400/1\*-IPQlOd46dlsutIbUq1Zcw.png](https://miro.medium.com/v2/resize:fit:1400/1*-IPQlOd46dlsutIbUq1Zcw.png)
6. LinkedIn. (2023). . <https://media.licdn.com/dms/image/D4D12AQHFJPGKuPjmDg/article-cover_imageshrink_600_2000/0/1685739097091?e=2147483647&v=beta&t=qXmPm8UIteILTcuCoLzOKYeTGJTo17uoHdt2z2e4dZ8>
7. “1\*5OQUNI-NR-S0GAZFIIt1RW.png (355×290).” [https://miro.medium.com/v2/resize:fit:710/1\*5OQunI-NR-S0gAZFIit1Rw.png](https://miro.medium.com/v2/resize:fit:710/1*5OQunI-NR-S0gAZFIit1Rw.png)
8. “2560px-Linear\_regression.svg\_wwqz1f3.png (2560×1693).” https://production-media.paperswithcode.com/methods/2560px-Linear\_regression.svg\_wwqz1f3.png
9. “5f6315207ab68b113cf57b1c\_Screenshot 2020-09-17 at 09.49.20.png (1246×828),” Sep. 17, 2020. <https://assets.website-files.com/5e6f9b297ef3941db2593ba1/5f6315207ab68b113cf57b1c_Screenshot%202020-09-17%20at%2009.49.20.png>
10. “1\*-IPQlOd46dlSUTIBUQ1ZCW.png (1723×1135).” [https://miro.medium.com/v2/resize:fit:3446/1\*-IPQlOd46dlsutIbUq1Zcw.png](https://miro.medium.com/v2/resize:fit:3446/1*-IPQlOd46dlsutIbUq1Zcw.png)
11. “Mango: Scalable hyperparameter tuning for AutoML,” Jul. 07, 2020. <https://community.arm.com/arm-research/b/articles/posts/scalable-hyperparameter-tuning-for-automl>
12. GeeksforGeeks, “Regression in machine learning,” GeeksforGeeks, <https://www.geeksforgeeks.org/regression-in-machine-learning/> (accessed Apr. 19, 2024).
13. S. Ray, “7 regression techniques you should know!,” Analytics Vidhya, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ (accessed Apr. 19, 2024).

# Appendix 1

The link for the Jupyter notebooks, visualization files and datasets can be found in the following GitHub link:

<https://github.com/Rishabh-Panchal/capstone-cra>

# Appendix 2

The link for the Jira is provided below:

<https://capstone-cra-bisi2024-grp3.atlassian.net/jira/software/projects/CP/boards/1/timeline>