



Machine Learning Model Application and Implementation in Project Management Owner Cost Predictions using SAS

Mohd Atir*

Mohd Atir, Kucha Chelan, New Delhi, India

*Corresponding author: Mohd Atir, Mohd Atir, Kucha Chelan, New Delhi, India.

Citation: Atir, M. (2025). Machine Learning Model Application and Implementation in Project Management Owner Cost Predictions using SAS. J. Robot. AI Comput. Appl. 1(2), 01-09.

Abstract

Owner cost prediction using machine learning techniques help in project cost allocation at the portfolio level. The application of ML algorithms using key factors like project size, project location, duration, extension, and so on turned out to improve the accuracy of the cost estimation by 14%. The final Mean absolute percentage error of the model turned out to be 25.16%. This research will serve as a foundation for other estimating factors. This paper will serve as a reference for application of machine learning models using historical data to predict the project owner cost for various capital-intensive projects.

Keywords: Cost Estimation, Machine Learning, SAS, Project Management

1. Introduction

Capital projects are cost intensive and require rigorous estimation of multiple factors. According to Enshassi et al, its successful realization depends on the accurate cost estimate which has a direct impact on the project profitability [1]. The objective of this paper is to have a comprehensive predictive capability to allocate the percentage of Project owner cost to support capital projects' decision making. This predictive capability is utilized through machine learning algorithms like Random Forest, Gradient Boosting, Polynomial Regression in SAS Viya platform. Macro-economic factors like inflation are utilized to increase the predictive capability of these machine learning algorithms.

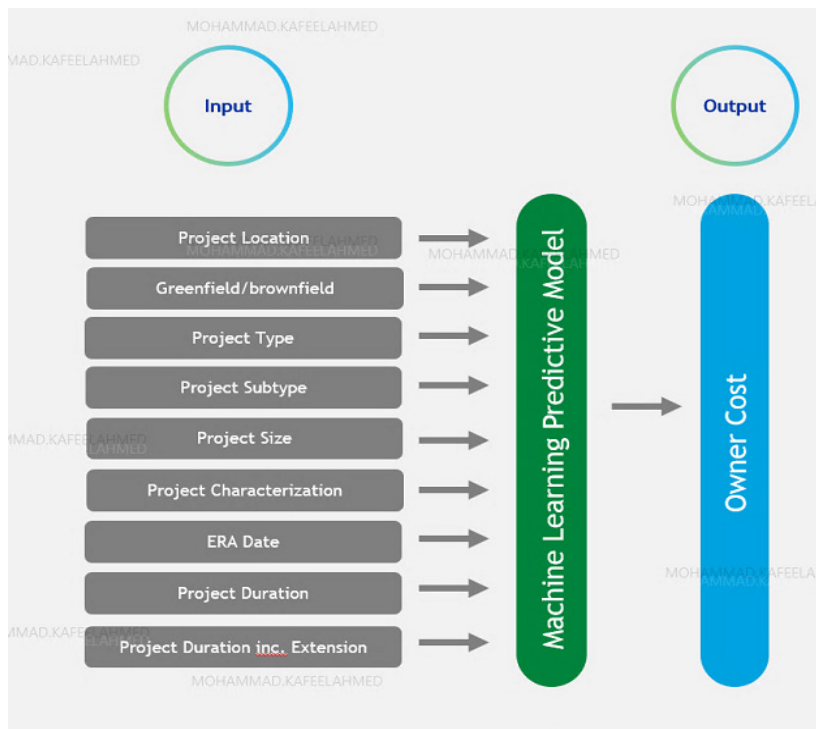
2. Literature Review

Cost estimating is the process of collecting and analyzing data [3] for any capital-intensive project. Bottom up estimating approach is often used by owners with detailed specifications of the parameters [3]. Back Propagation Neural networks and SVMs had been used as an approach for cost estimations based on product lifecycle [4]. This paper uses the machine learning approach to predict the owner cost, which has not been a standard in cost estimations of the capital-intensive projects.

3. Benefits

- Optimize Owner cost allocation at the portfolio level.
- Serve as foundation for advanced analytics for other estimating factors such as change orders and contingency.

4. Model diagram

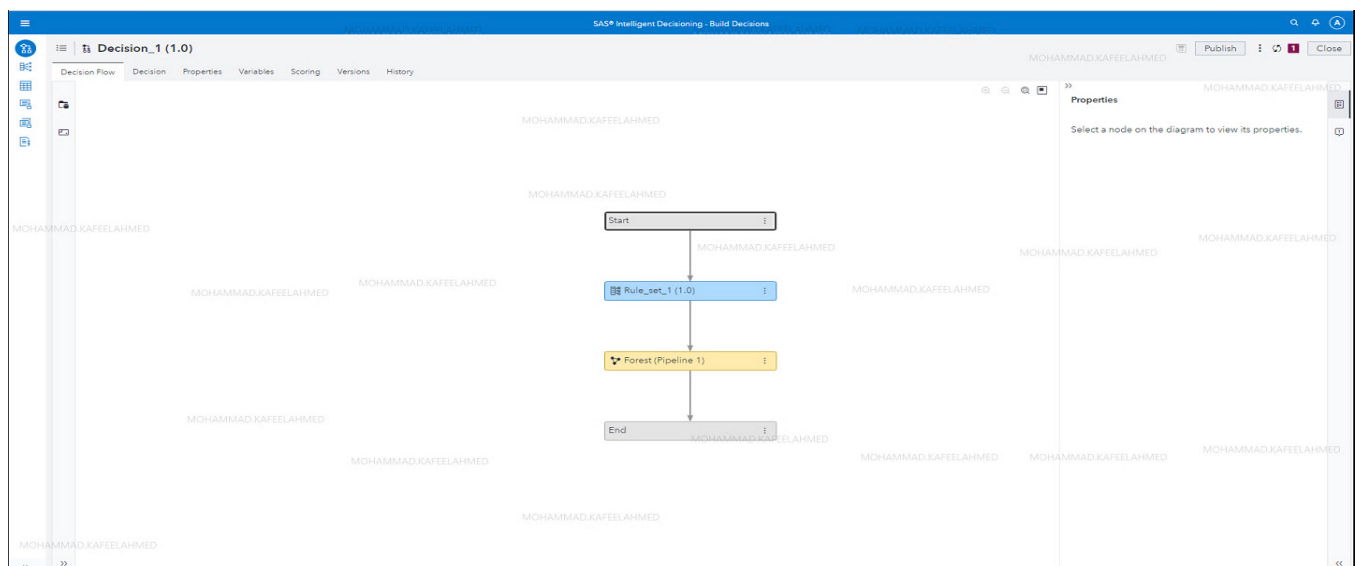


Following important features had been used for developing the model:

- Project Size – total cost of the project refers to its size. Since it was an historical data for 15 years therefore, inflation adjustment had been applied as the regional level using inflation rate from world bank.
- Project Characterization – it refers to the type of the project based on its characters being A, B, or C.
- Project Type and Project Subtype refers to the type of the projects being construction, exploration, telecommunication, and various other types and subtypes.
- Project duration and extension refers to the total time the project took to be completed and if there is any delay in the completion. The duration and extension are reflected in months.
- Other factors like Greenfield/brownfield, ERA date, Project Location were used in the model development as independent factors.

5. Build Decisions Development Process Flow

SAS Decision Manager helps organizations manage data, business rules, analytical models, and optimization techniques. Rule management, model management, and data management are integrated into a consistent interface for easier accessibility.



The decision flow shown above in the screenshot is the SAS Viya pipeline for the deployment of the random forest model that was shortlisted as the final model.

This decision is published on Micro Analytic Server to be consumed by the web service.

The inputs will be passed to the webservice, and this decision flow will create new features from those inputs, and run the scoring model iteration on those features, and provide the predicted owner cost value.

6. Rule Sets were used to Develop New Features to Pass as Inputs to the Forest Model Pipeline

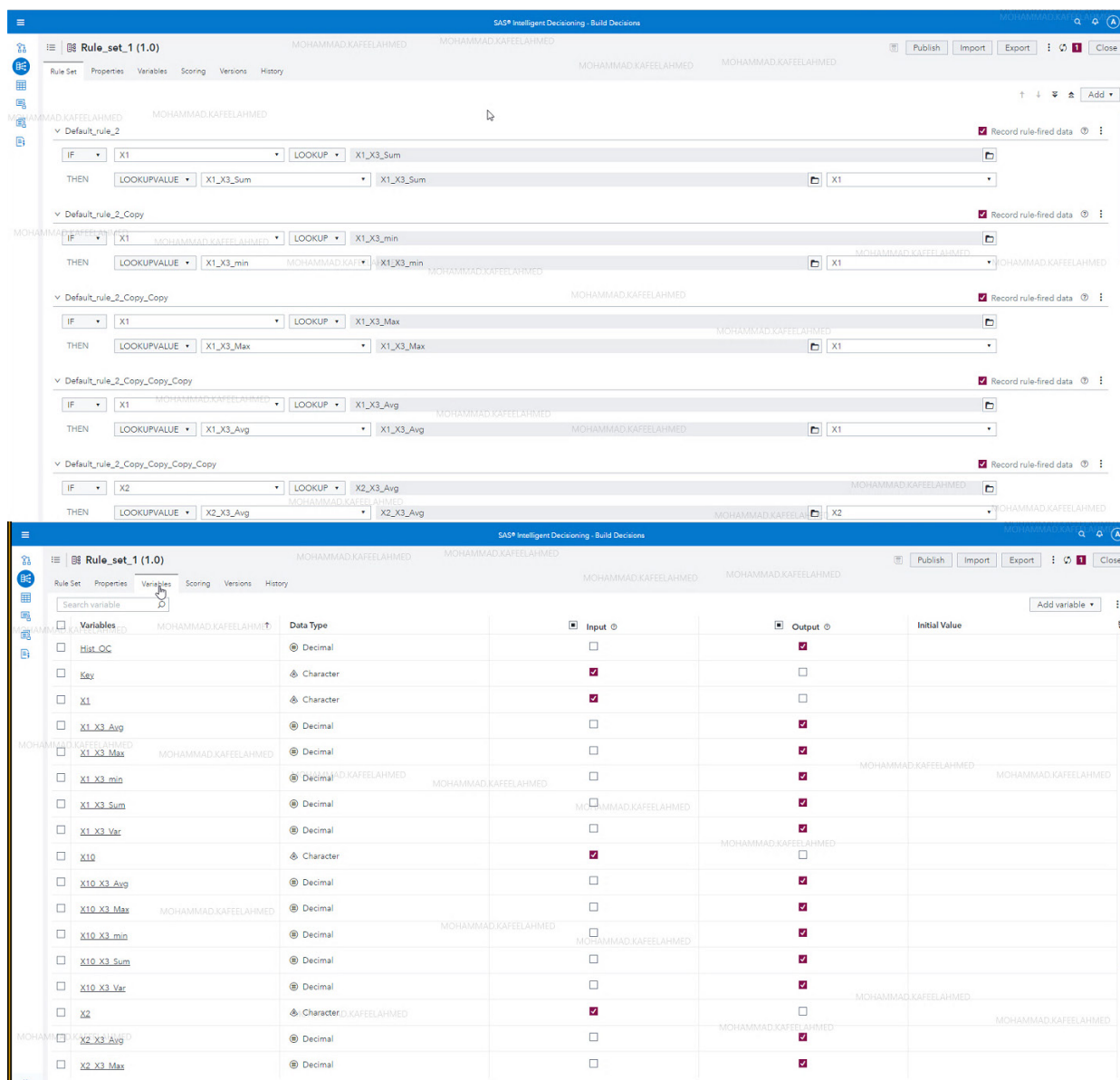
A rule specifies conditions to be evaluated and actions to be taken if those conditions are satisfied. Rules are grouped together into rule sets. Rule sets are logical collections of rules that are grouped together because of interactions or dependencies between the rules or because they are processed together when they are published.

Most rules correspond to this form:

if condition_expressions then action_expressions

For example, suppose you have the following rule:

if customer_debt > customer_assets then app_status = 'Decline'



Variables	Data Type	Input	Output	Initial Value
Hist_OC	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
Key	Character	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
X1	Character	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
X1_X3_Avg	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X1_X3_Max	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X1_X3_min	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X1_X3_Sum	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X1_X3_Var	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X10	Character	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
X10_X3_Avg	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X10_X3_Max	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X10_X3_min	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X10_X3_Sum	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X10_X3_Var	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X2	Character	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
X2_X3_Avg	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
X2_X3_Max	Decimal	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

7. Lookup Tables were used to Develop Rules for the Input Variables

SAS Decision Manager provides the ability to import lookup tables and reference them from rules. Lookup tables are tables of key-value pairs. For example, we can use a lookup table to retrieve a part name based on the part code number. Or use a lookup table to retrieve the full name for a country based on its abbreviation.

We can import lookup data from comma-separated-values (CSV) files such as those created by Microsoft Excel into lookup tables in SAS Decision Manager. We can re-import updated CSV files as needed to refresh the lookup tables.

Name	Description	Location	Modified By	Date Modified
Subject Level			sas.decisions	Dec 19, 2019 03:28 PM
Treatment Channels			sas.treatmentDefinitions	Dec 19, 2019 03:28 PM
X1_X3_Sum		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:51 AM
X1_X3_min		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:52 AM
X1_X3_Max		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:52 AM
X1_X3_Avg		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:53 AM
X1_X3_Var		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:59 AM
X2_X3_Sum		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 09:59 AM
X2_X3_min		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:00 AM
X2_X3_Max		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:00 AM
X2_X3_Avg		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:01 AM
X2_X3_Var		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:01 AM
X10_X3_Sum		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:02 AM
X10_X3_min		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:06 AM
X10_X3_Max		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:06 AM
X10_X3_Avg		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:07 AM
X10_X3_Var		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:07 AM
Hist_OC		/Users/kafeelma/My Folder	kafeelma	Mar 10, 2022 10:15 AM

The inflation rate was used to calculate the Inflation adjusted values of the target variable and input variable over the years.

Inflation adjusted calculation was done to the Owner cost and project size over the years as the data period ranges from 2005 to 2018. Inflation had a major role in the deciding the owner cost and project size from the business perspective.

8. New Features were used in the Model

There is no well-defined procedure for feature engineering as it primarily depends on domain knowledge, and trial and error on the available data [2]. The algorithms for dimensionality reduction like PCA could play a crucial role in feature engineering, but it is not advisable to apply them without proper data exploration and feature engineering.

Based on Project independent factors, the following features were defined:

- Number of Days Delay in the Project Completion**

This feature is determined based on Days Delay for project completion, which is calculated as

$$D_c = p_c - t_c$$

D_c refers to the number of days delay in the project completion for each vendor c , which was derived from subtracting t from p . t refers to the day on which project was supposed to complete and p refers to the day the project was actually completed.

For c^{th} project the summation of D is taken for the i^{th} record, from first project.

$$D_c = \sum_{k=1}^i D_{c,k}$$

- Number of Days Since Last Project**

The difference of days is taken for each vendor c since its last project $t-1$.

$$J_c = t_c - (t_c - 1)$$

J for each vendor c is calculated as the difference in the days from its current project start date till the previous project completion date $(t-1)$.

- **Flag Referring to Previous Project Delay or Non-Delay**

This feature indicates the Delay of the previous project $i-1$ for the vendor c .

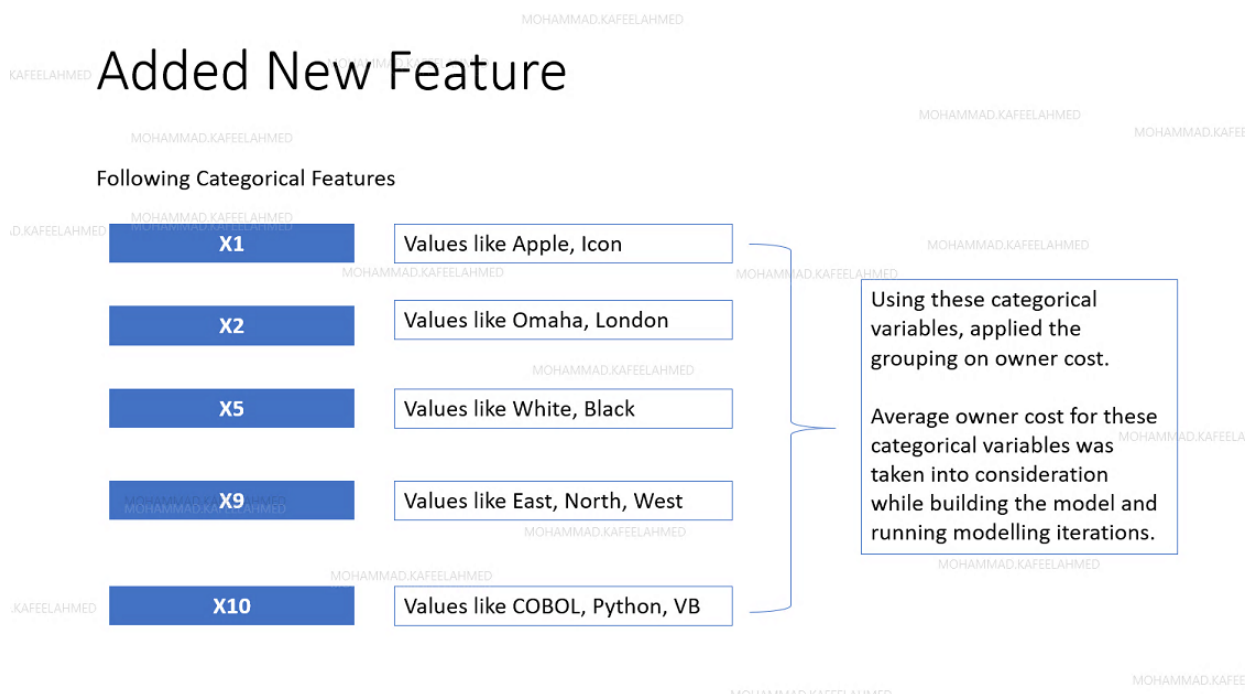
$$F_c = F_{c,(i-1)}$$

It is a binary flag taken from the value of F from the previous project $(i-1)$. i^{th} project is the current project for the vendor c .

- **Balance Due**

Balance due as parameter is calculated as the difference between Invoice amount and the paid amount as calculated below:

$$A_c - M_c$$



The features like project characterization, type, subtype, and so on are mentioned as $X1, X2, X3, \dots, Xn$

The categorical factors (mentioned $X1-X10$) used in the model are used for feature engineering, and new features were derived from them.

New features in the model can be created through following SASmacro:

```
%macro mean_encoding(dataset,var,target);
proc sql;
  create table mean_table as
  select distinct(&var) as gr, mean(&target) As mean_&var,
  std(&target) As std_&var,
  median(&target) As median_&var,
  max(&target) as max_&var,
  min(&target) as min_&var
```

group by gr;

```
select d.*, m.mean &var, m.std &var, m.median &var, max &var, min &var
```

```
left join mean table as m
```

```
on &var=m.gr;
```

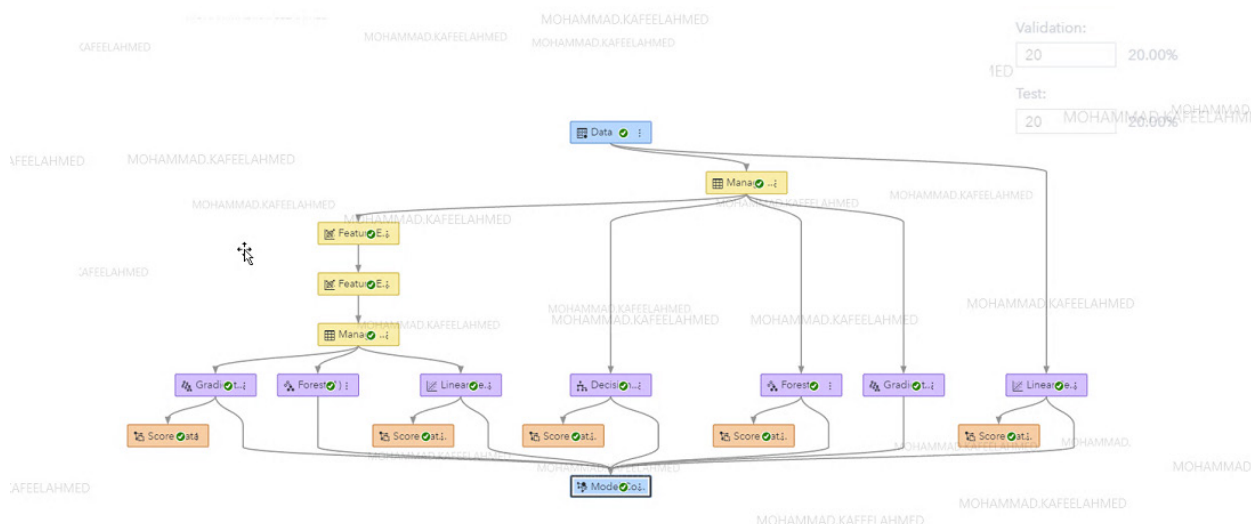
quit;

```
%mend;
```

```
%mean encoding(class,age,height);
```

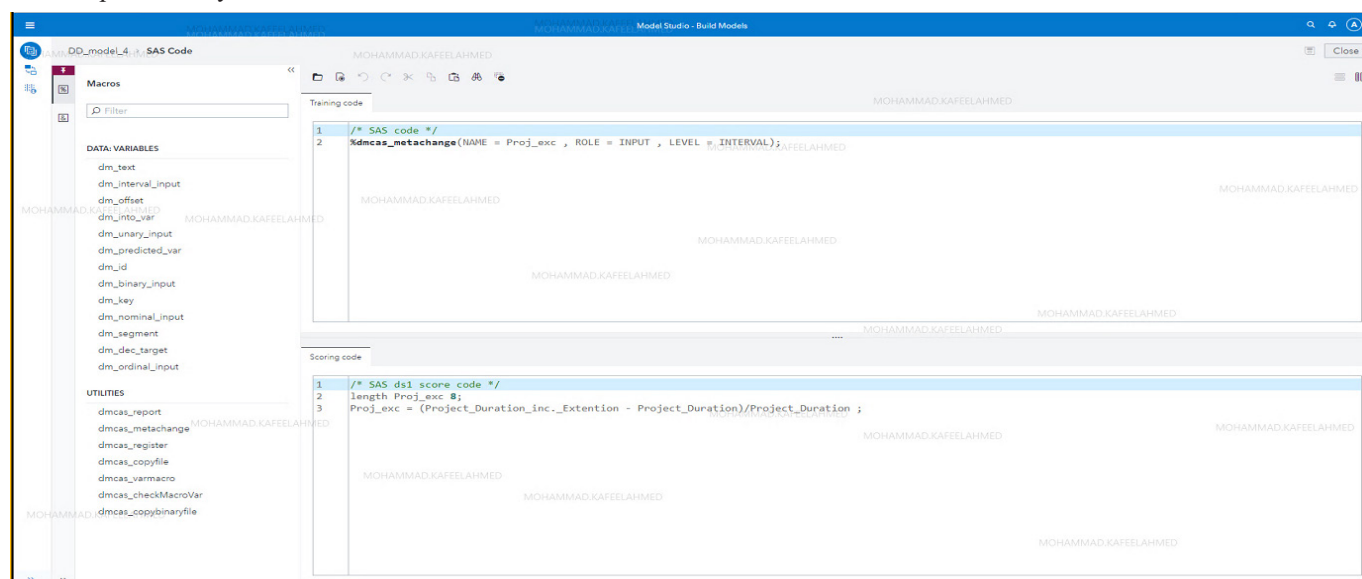
6. SAS Viya Pipeline

SAS Viya Pipeline models were used as follows in the below figure:



7. SAS Code Node

In built SAS Code node in SAS Viya Pipeline was also used to derive calculated input features for the model. The below figure gives an example of the syntax used to derive new features:



8. Brief About Machine Learning Algorithms Applied in Project Owner Cost use case Linear Regression

OLS is the method with which linear regression is performed. The square of the difference from the mean is taken for every data point, and the summed loss function is to be minimized.[8]

$$l = \sum_{i=1}^n (y_i - \bar{y})^2$$

Polynomial

Polynomial regression is a modification of linear regression where the existing features are mapped to a polynomial form. The problem is still a linear regression problem, but the input vector is now mapped to a higher dimensional vector which serves as a pseudo-input vector of sorts.[8]

$$x = (x_0, x_1) \rightarrow x' = (x_0, x_0^2, x_1, x_1^2, x_0x_1)$$

Stepwise

Stepwise regression or spline regression helps us fit a piece wise function to the data. It is usually used with linear models, but it can be generalized to higher degrees as well. The regression equation takes the form of

$$y = ax + b(x - x_0)H_{\alpha} + c \quad \text{or} \quad y = ax + b(x - x_0)H_{\alpha} + c$$

where H_{α} is the shifted Heaviside step function, having its discontinuity at α . [8]

Decision Trees

Decision tree works by successively splitting the dataset into small segments until the target variable are the same or until the dataset can no longer be split. It's a greedy algorithm which make the best decision at the given time without concern for the global optimality[5].

The concept behind decision tree is straightforward. If the address is "myEmployer.com", it will classify it to "Email to read when bored". Then if the email contains the word "hockey", this email will be classified as "Email from friends". Otherwise, it will be identified as "Spam: don't read".

Random Forest

Random forest is a bagging algorithm that takes the trees that are constructed with the injection of randomness, and that's why this algorithm is called 'random forest'. It takes into consideration many input variables without overfitting. The convergence in the random forest depends on the strong features, and not by the noise in the variables present.[5] It is a bagging-based algorithm developed by Leo Breiman coined from 'bootstrap aggregating'. It uses multiple classifiers' voting to assign a class or label to the training samples.[6]

Gradient Boosting

Boosting based algorithms like GBM also use multiple classifiers to assign a class to the training set, but they also assign weights to the training samples, and increase these weights as they move into the next classifiers. To explain it further, it will assign equal weights to all the training samples in the first classifier, and as they move into the next classifier it assigns more weights to the training samples that have been incorrectly classified. [7]

9. How To Call SAS MAS REST API

In order to deploy the model, the REST APIs were used to be consumed through SAS Micro analytic server. Following is the process for the ML model deployment for its utilization by the business users.

For SAS Administrator

To be able to use SAS MAS REST API we need to register a new client on SAS server

1. login to SAS server from the terminal with SAS admin user

1. get the Consul-Token by opening the following file (not recommend using cat, instead use export):

```
/opt/sas/viya/config/etc/SASSecurityCertificateFramework/tokens/consul/default/client.token
```

3. Run the following command to generate a client registration token using the Consul-Token

```
curl -k -X POST "https://webpage/SASLgon/oauth/clients/consul?callback=false&serviceId={client_id}" -H "X-Consul-
```

Token: {Consul-Token}"

4. Run the following command to register new client using the client registration token:

```
curl -X POST "http:// weblink/SASLogon/oauth/clients" -H "Content-Type: application/json" -H "Authorization: Bearer {client_token}" -d '{"client_id": "{client_id}", "client_secret": "{client_password}", "scope": ["openid"],"authorized_grant_types": ["client_credentials"], "authorities": "{client_group}", "access_token_validity": 86400}'
```

For Developers

5. Run the following command to generate OAUTH access token:

```
curl -k -X POST "https:// weblink/SASLogon/oauth/token" -H "Content-Type: application/x-www-form-urlencoded" -d "grant_type=client_credentials" -u "{client_id}:{password}"
```

6. Call SAS MAS REST API:


```
curl --location --request POST 'https:// weblink/microanalyticScore/modules/owner_cost_prediction/steps/score' \
--header 'Authorization: Bearer {access_token}' \ --header 'Content-Type: application/json' \ -d '{ "inputs": [ { "name": "X1", "value": "0" }, { "name": "X2", "value": "0" }, { "name": "X5", "value": "0" }, { "name": "X6", "value": 0 }, { "name": "X7", "value": 0 }, { "name": "X8", "value": 0 }, { "name": "X9", "value": "0" }, { "name": "X10", "value": "0" }, { "name": "Key", "value": "0" } ] }'
```

After its deployment, the business users would just have to input the values of the independent variables, and the predicted owner cost of the project will be the output of the web app.

Results and Discussion

The error rate using these Machine Learning algorithms in built in SAS Viya was quite high as it was going up to 40% on the holdout data. After applying feature engineering discussed in section 8 of this paper, the mean absolute percentage error using these algorithms reduced to ~25.16%.

Final Root mean Squared Logarithmic error from this model is as follows:

OC_model_whole_data > "Model Comparison" Results				
Node				
Assessment				
Model Comparison				
Champion	Name	Algorithm Name	Root Mean Squared Logarithmic Error	
	Forest	Forest	0.5881	
	Decision Tree	Decision Tree	0.6948	
	Linear Regression (1)	Linear Regression	0.7599	
	Gradient Boosting (1)	Gradient Boosting	0.7696	
	Gradient Boosting	Gradient Boosting	0.7862	
	Forest (1)	Forest	0.9585	
	Linear Regression	Linear Regression	1.9433	

References

1. Enshassi, A., Mohamed, S., & Madi, I. (2015). Cost estimation practice in the Gaza Strip: A case study. IUG Journal of Natural Studies, 15(2).
2. [Khurana, U., Samulowitz, H., & Turaga, D. \(2018, April\). Feature engineering for predictive modeling using reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence \(Vol. 32, No. 1\).](#)
3. Choon, T. T., & Ali, K. N. (2008). A review of potential areas of construction cost estimating and identification of research gaps. Journal Alam Bina, 11(2), 61-72.
4. [Yeh, T. H., & Deng, S. \(2012\). Application of machine learning methods to cost estimation of product life cycle. International Journal of Computer Integrated Manufacturing, 25\(4-5\), 340-352.](#)
5. Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research, 13, 1063-1095.
6. [Breiman, L. \(2001\). Random forests. Machine learning, 45, 5-32.](#)
7. [Machová, K., Puszta, M., Barčák, F., & Bednár, P. \(2006\). A comparison of the bagging and the boosting methods using the decision trees classifiers. Computer Science and Information Systems, 3\(2\), 57-72.](#)
8. [Ostertagová, E. \(2012\). Modelling using polynomial regression. Procedia engineering, 48, 500-506.](#)

Copyright: ©2025 Mohd Atir. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.