Linear Statistical Modeling Methods with SAS

Introduction to SAS Viya Model Studio

Xuemao Zhang
East Stroudsburg University

April 17, 2024

Outline

- Overview of SAS Viya Model Studio
- An Example of Supervised Learning

Overview of Model Studio

SAS Viya Model Studio is a web-based platform which includes a suite of integrated data mining tools to facilitate end-to-end data mining analysis.

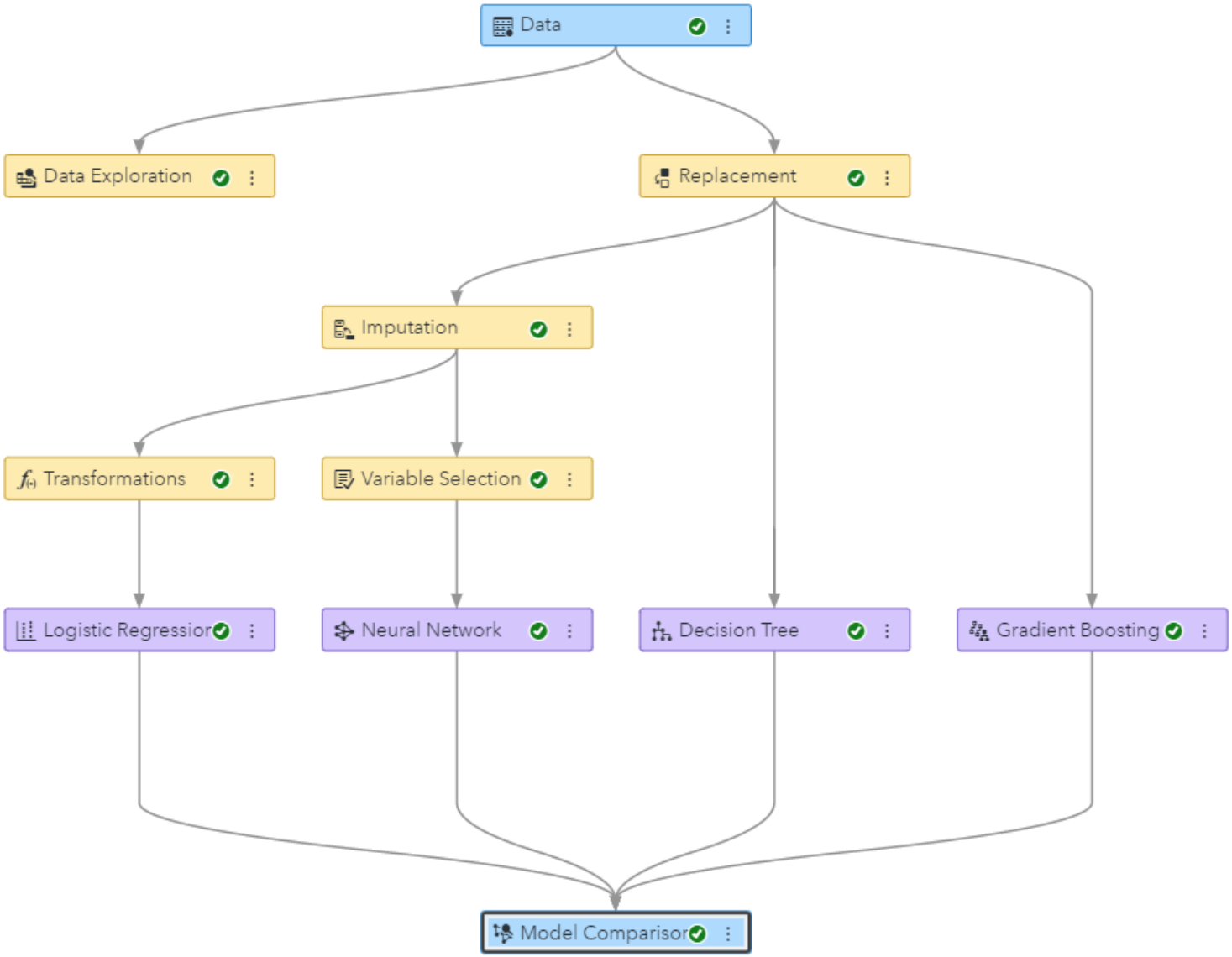
- Model Studio contains the following SAS solutions:
  - SAS Visual Forecasting
  - **SAS Visual Data Mining and Machine Learning**
  - SAS Visual Text Analytics
- We check the main functionality of SAS Visual Data Mining and Machine Learning
  - SAS Visual Data Mining and Machine Learning: User's Guide

Overview of Model Studio

- SAS Visual Data Mining and Machine Learning Features List
  - Visual interface for the entire analytical life cycle process.
  - Drag-and-drop interactive interface requires no coding, though coding is an option.
  - Supports automated code creation at each node in the pipeline.
  - Choose best practice templates (basic, intermediate or advanced) to get started quickly with machine learning tasks or take advantage of our automated modeling process.
  - Embed open source code (such as R and Python) within an analysis, and call open source algorithms within Model Studio.
  - Automatically generates SAS DATA step code for model scoring; Applies scoring logic to training, holdout data and new data.
  - $\vdots$

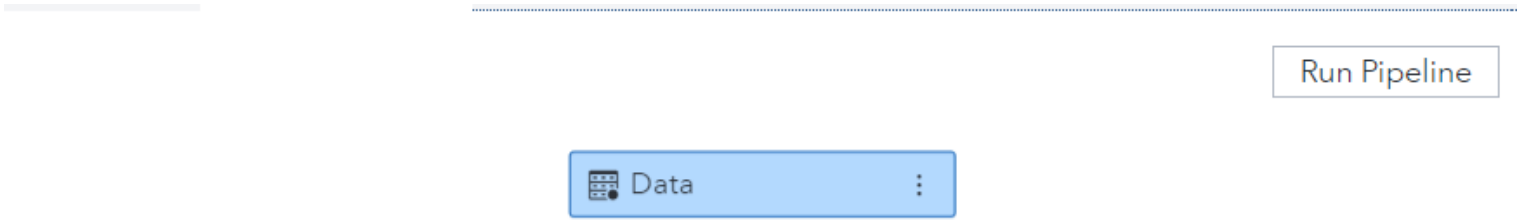
Overview of Model Studio

- data mining pipeline



An Example

- In Lecture 16, we considered the `Auto` data and regarded `mpg` as the response, and used `PROC GLMSELECT` select a best multiple linear regression model.
- Our data is clean, and the only data cleaning step we did before is converting the two variables `origin` (3 categories) and `cylinders`(5 categories) to categorical.
  - SAS Viya just recognizes them as nominal or categorical variables.
  - We ignored the predictor `year`. You may add it as a categorical variable to the model.
- Our data mining pipeline is like this



An Example

- **Step 1**: **Create a new project**
  - Type: `Data Mining and Machine Learning`
  - Template: `Blank template`
  - Data: Find the `Auto` data we uploaded before
- Click `Advanced` to have a quick look of advanced settings
  - Advisory Options: Rule to determine if a variable is categorical
  - Partition Data: Partitioning data into `Training`, `Validation` and `Test` data
    - ⋆ Let choose method `Simple random` (default is `stratify`) and 70%,30% and 0% for the `Training`, `Validation` and `Test` data.
    - ⋆ **Note**: Partition Data Create partition variable Note: These settings are active only when a partition variable is not set within the data. Using a data source with a pre-defined partition variable or manually selecting a partition variable will **override** these settings.
  - Event-Based Sampling: it is for rare-event response. Do not enable it.
  - Node Configuration: This setting is useful when we use the Open Source Code node with the language set to Python.

An Example

- We'll see a warning message `You must assign a variable with the role of Target in order to run a pipeline.`
- **Step 2**: In the `Data` tab which is selected by default, **assign the** `Role` of the response variable `mpg` to `Target`.
  - SAS Viya can tell that the variable is `interval` or numerical.
  - `name` is assigned as `ID`
- We can reject some variables in the `Data` tab
  - Let's reject the variable `year` (suppose `year` does not make any difference). You may keep it and verify if it is significant. But let's reject it for now.

An Example

- **Step 3**: We may **update the data partition** by clicking `Project Settings` the in the upper-right corner.
  - The partition settings can be edited only if no pipelines in the project have been run.
  - After the first pipeline is run, the partition tables are created for the project, and the partition settings cannot be changed.
  - The `Rules` options can be used to change the selection statistic and partition data set that determine the champion model during Model Comparison. Statistics can be selected for categorical and numerical response (targets).
  - Click `Save` to save the new partition settings.

An Example

- **Step 4**: **Check the data mining pipeline**. Click the `Pipelines` tab. We currently have a single pipeline in our project, which is called Pipeline 1. It currently contains only a **Data** source node.
  - you may change the name of the pipeline.
  - right-click the node and then select `Run`.
  - The green check mark indicates that the node ran successfully, without an error, and the data has been partitioned.
- After the Data node has been run, we cannot change the partitioning, event-based sampling, project metadata, project properties, or the target variable.
  - However, we can change variable metadata with the Manage Variables node or through the *Data* tab
- Log file: click `Settings` in the top right corner. And then select Project logs.
  - From the available logs, select Log for Project Partitioning, and then click Open.
  - The log file can be downloaded for record keeping.

An Example

- **Step 5**: **Data Exploration**.
  - Click the *Pipeline1* tab. In the pipeline, right-click the *Data* node and select *Add child node -> Miscellaneous -> Data Exploration*. The *Data Exploration* node will be added to the pipeline and and connected to the *Data* node.
  - Another way to bring a node into a pipeline is to click the **Nodes icon** in the left panel and drag one of the listed nodes on top of a node that is already in the pipeline.
- Run the *Data Exploration* node.
- When the run is complete, open the Data Exploration *results*.
  - Data Partition Summary
  - Important Inputs
  - Class Variable Summaries
  - Class Variable Distributions
  - Interval Variable Moments
    - ⋆ Skewness normality range is $(-1, 1)$ or acceptable $(-2, 2)$
    - ⋆ Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution; normality range $(-2, 2)$ or acceptable $(-7, 7)$
    - ⋆ If a variable has high Skewness or Kurtosis, transformations will be needed.
  - $\vdots$

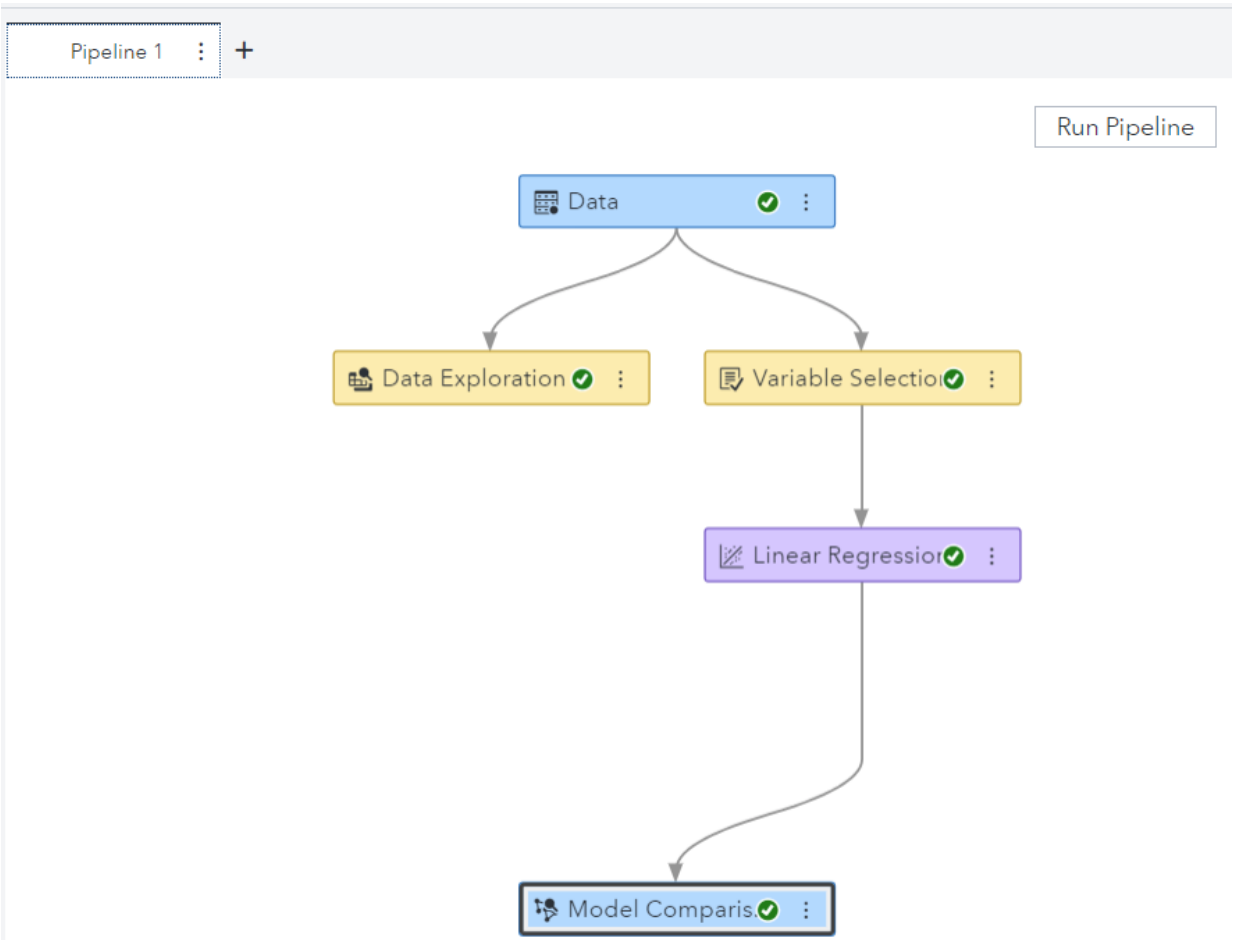
An Example

- **Step 6**: **Variable Selection**. We add a *Variable Selection* as a child node to *Data* node.
  - Let's choose methods `Unsupervised Selection`, `Fast Supervised Selection`, and `Linear Regression Selection`
  - Run the node and check the results.
  - `Unsupervised Selection` is done by the `VARREDUCE Procedure`

- **Step 7**: Add a **Linear Regression** node to the pipeline.
  - It is a child node of the `Variable Selection` node.

An Example

- **Step 8**: **Run the pipeline and check the results**.
  - Check the mode fit and selected predictors from the `Linear Regression` node.
  - There's only one model in the pipeline, we can look at the performance of the model through the **Model Comparison** node.



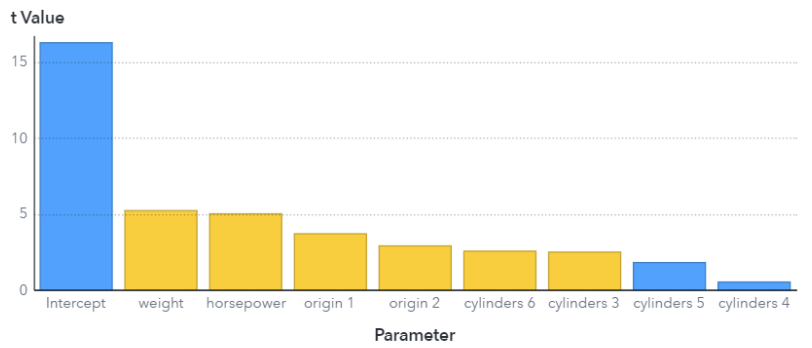
An Example

Auto_project > **"Linear Regression" Results**

Summary | Output Data

Node | Assessment | Model Interpretability

t Values by Parameter



Parameter Estimates

Effect	Parameter	t Value	Sign	Estimate
Intercept	Intercept	16.2966	+	45.4249
weight	weight	5.2733	-	-0.0040
horsepower	horsepower	5.0642	-	-0.0707
origin	origin 1	3.7540	-	-2.8339
origin	origin 2	2.9587	-	-2.4215
cylinders	cylinders 6	2.6135	-	-2.6642
cylinders	cylinders 3	2.5512	-	-7.8309
cylinders	cylinders 5	1.8649	+	5.8523

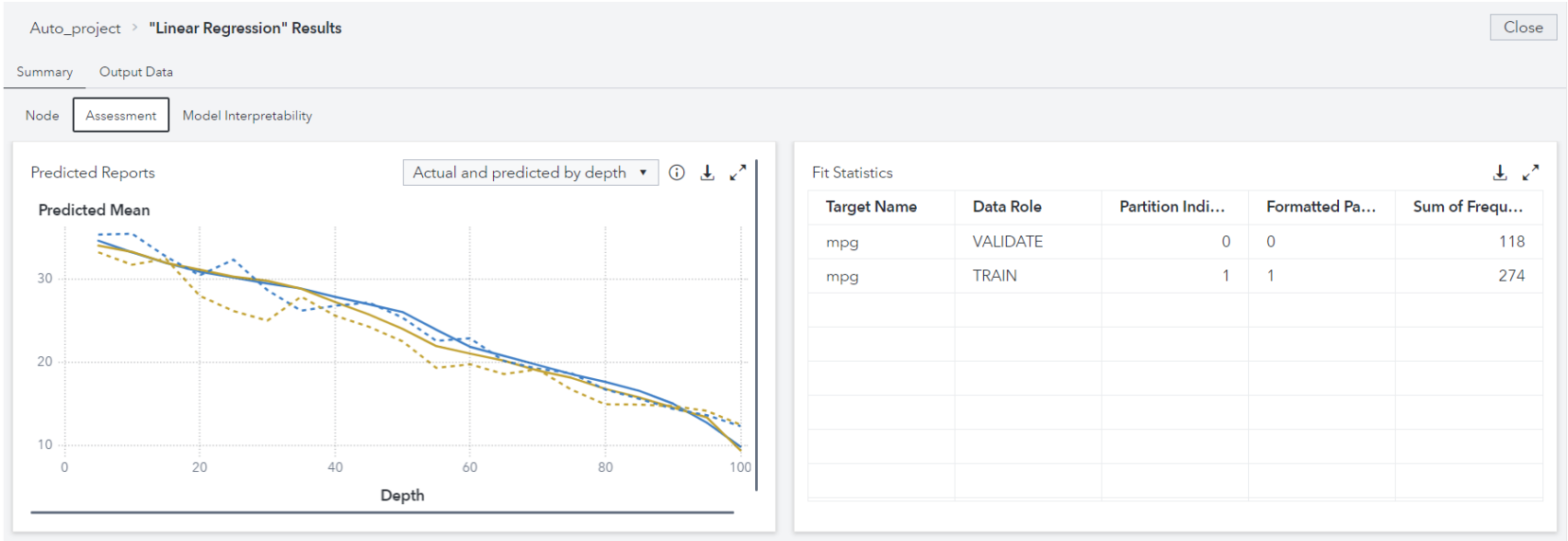
Selection Summary

Step	Effect Entered	Number of E...	SBC	Optimal SBC
0	Intercept	1	1,147.2864	0
1	weight	2	827.9212	0
2	horsepower	3	814.7946	

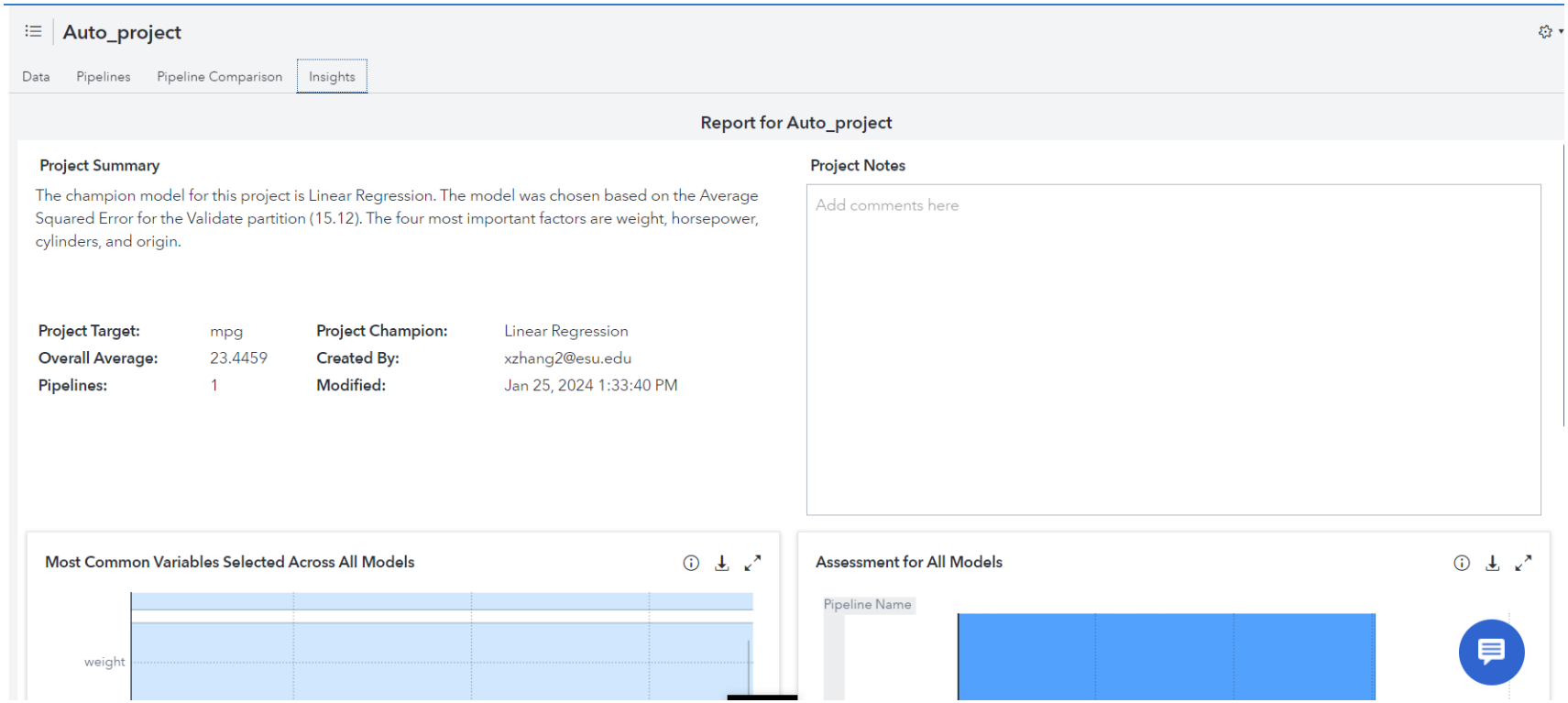
Regression Fit Statistics

Statistic	Description	Value
RMSE	Root MSE	
RSQUARE	R-Square	
ADJRSQ	Adj R-Sq	0.7566

An Example


Auto_project > "Linear Regression" Results
Close
Summary
Output Data
Node
Assessment
Model Interpretability
Predicted Reports
Actual and predicted by depth
Predicted Mean
30
20
10
0
20
40
60
80
100
Depth
Fit Statistics
Target Name | Data Role | Partition Indi... | Formatted Pa... | Sum of Frequ...
mpg | VALIDATE | 0 | 0 | 118
mpg | TRAIN | 1 | 1 | 274

An Example



- Based on the model, we can then apply the model to the complete data and use the estimated regression equation for predictions. We continue in the next lecture.

License