Archetype

User Guide

V1.4

JAYWING

Contact information

Primary points of contact

Helpdesk:

Email: operations.helpdesk@jaywing.com

Locations

Leeds

Globe Point
1 Globe Road
Leeds
LS11 5FD

Sheffield

32 Eyre St
Sheffield
S1 4QZ

Contents

1 Introduction

Archetype is Jaywing's browser-based modelling software that delivers substantial improvements in risk model performance through models that are fully-controlled and explainable.

- Enables a significant reduction in bad debt for the same level of customer acquisition.
- Equally applicable to combating fraud - reducing the data and processing costs associated with fraud detection.
- Typically delivers improvement to Gini statistic of 5-10% on a like-for-like methodology basis vs logistic regression methods.
- Meets the higher bar of explainability the ICO requires of lenders: interpretability by design.
- Builds models in hours not weeks and enables rapid rebuild and re-calibration without having to re-visit a lengthy model governance process.

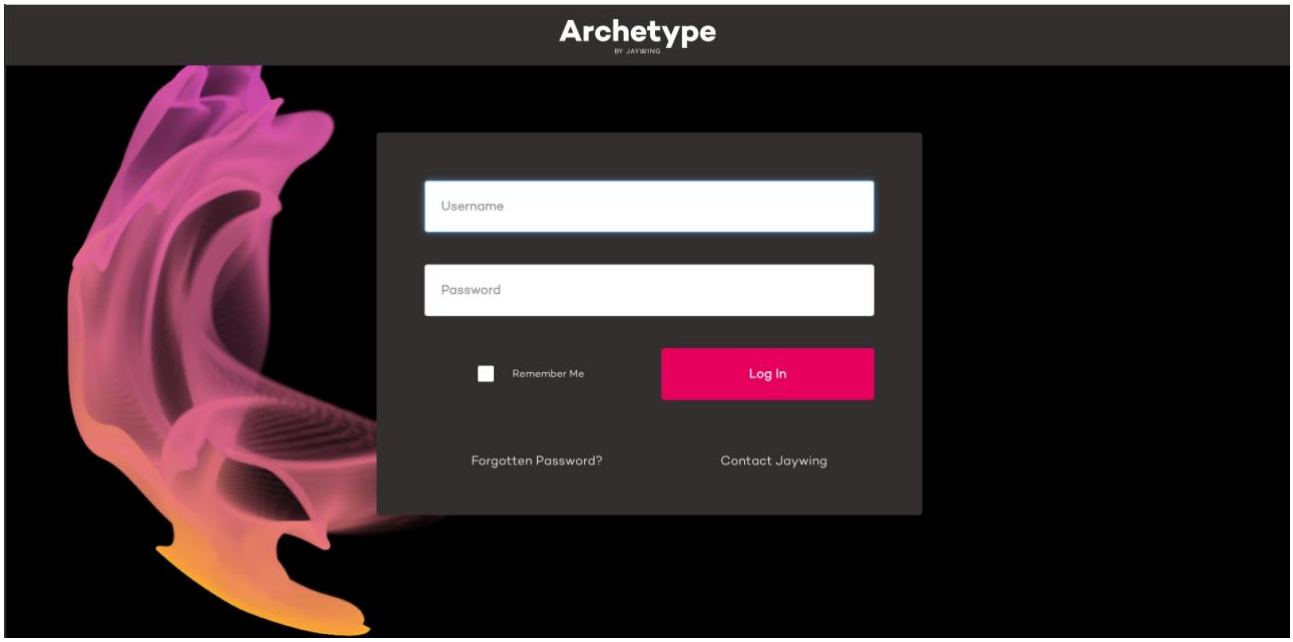
This document provides a step-by-step user guide of a modelling process using Archetype.

2 Log In

Archetype is a browser-tool. To access it, visit https://archetype.jaywing.com/login/.

User credentials (username and initial password) will be provided by the Jaywing IT team.

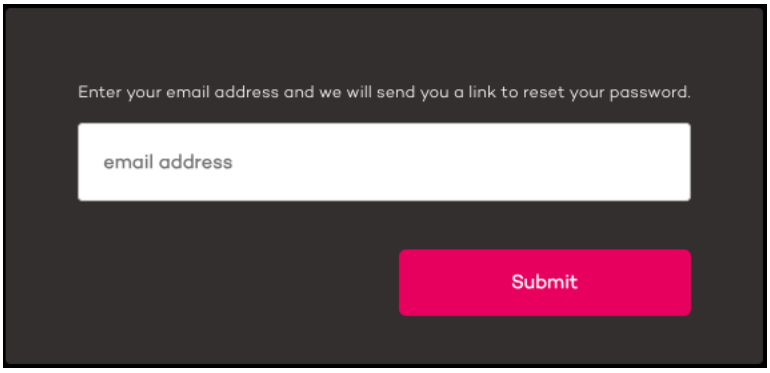
To reset your password or recover a forgotten one, click on "Forgotten Password?" to initiate the password recovery process.



Forgotten Passwod

Click on "Forgotten Password?" or visit: https://archetype.jaywing.com/login/forgotten-password

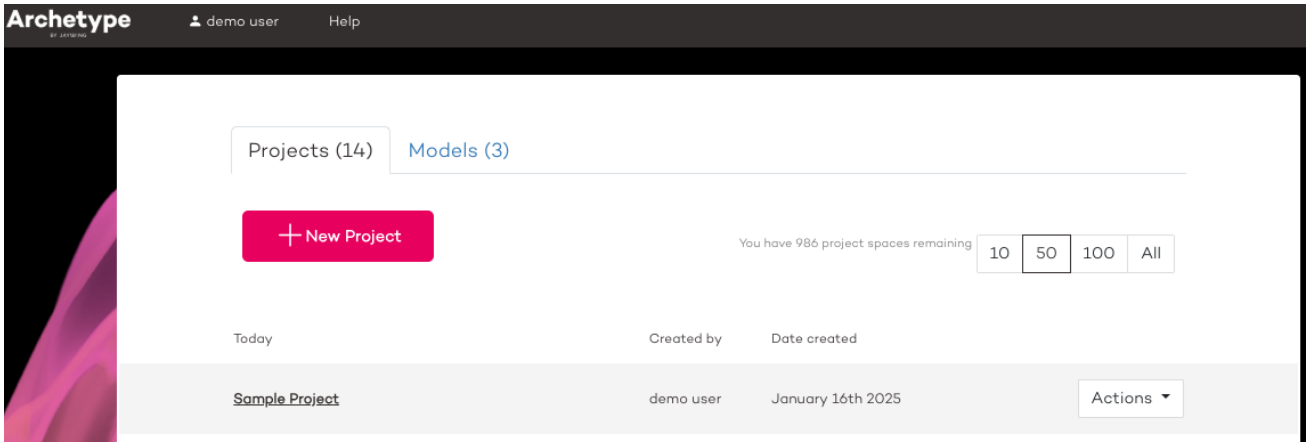
Provide your registered email address and click Submit.



3 Projects

3.1 Projects List Tab

After logging in, users will land on the page shown below:

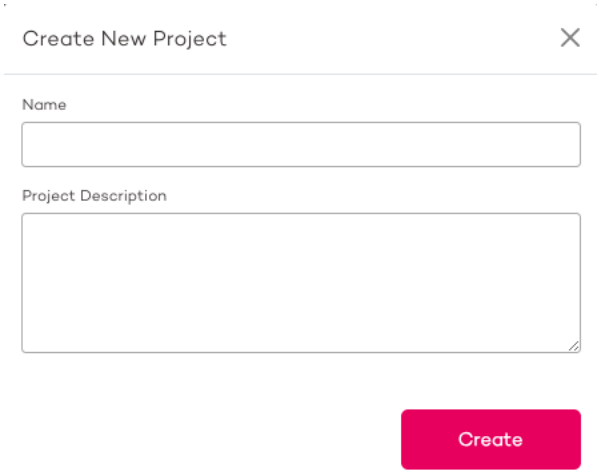


The page is divided into two tabs “Projects” and “Models”

Projects	Personal model development projects, as well as projects shared for collaboration. Typically, these are works in progress.
Models	Finalized, published models that are immutable, as they are prepared for deployment.

3.1.1 Create New Project

To initiate a new modelling process, select “Create New Project” button and fill in Name and short description (optional but recommended).



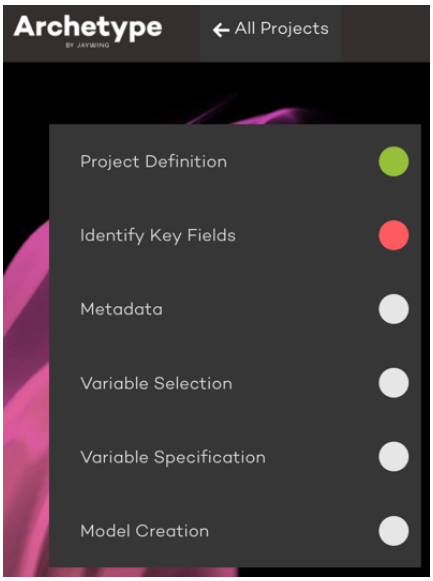
3.2 Modelling Stages

Archetype offers a standardised modelling process, ensuring consistency across models developed by different users and teams.

The process is divided into six distinct stages, which are displayed in the left sidebar of the interface. This structured layout provides a clear and organized workflow.

Additionally, Archetype uses a traffic light system to give users a quick visual summary of each stage's completion status.

See below table for a brief explanation of these stages and their corresponding indicators:



Modelling Stages

1	Project Definition	Defining project name, description and uploading the development data.
2	Identify Key Fields	Identifying key fields in the uploaded data, such as the unique ID, target values, train/validation split, and optionally record weights.
3	Metadata	Variable labels (descriptions), default values, and the treatment of variables (categorical or continuous) can either be defined manually or through file upload (from previous projects). If not provided, Archetype will make automated decisions, which can be adjusted later.
4	Variable Selection	To perform variable selection (manual or automated), several key statistics are calculated to help assess the importance or relevance of each feature in a model.
5	Variable Specification	Variable review supported by Partial Dependency Plots, with the optional application of monotonicity constraints, rank ordering, and the creation of broader bins (coarse classing).
6	Model Creation	Defining the model architecture and training the model.

Traffic Lights (Progress Lights)

Green	Stage completed
Red	Incomplete tasks.
Bule (flashing)	Active data processing (user should wait).

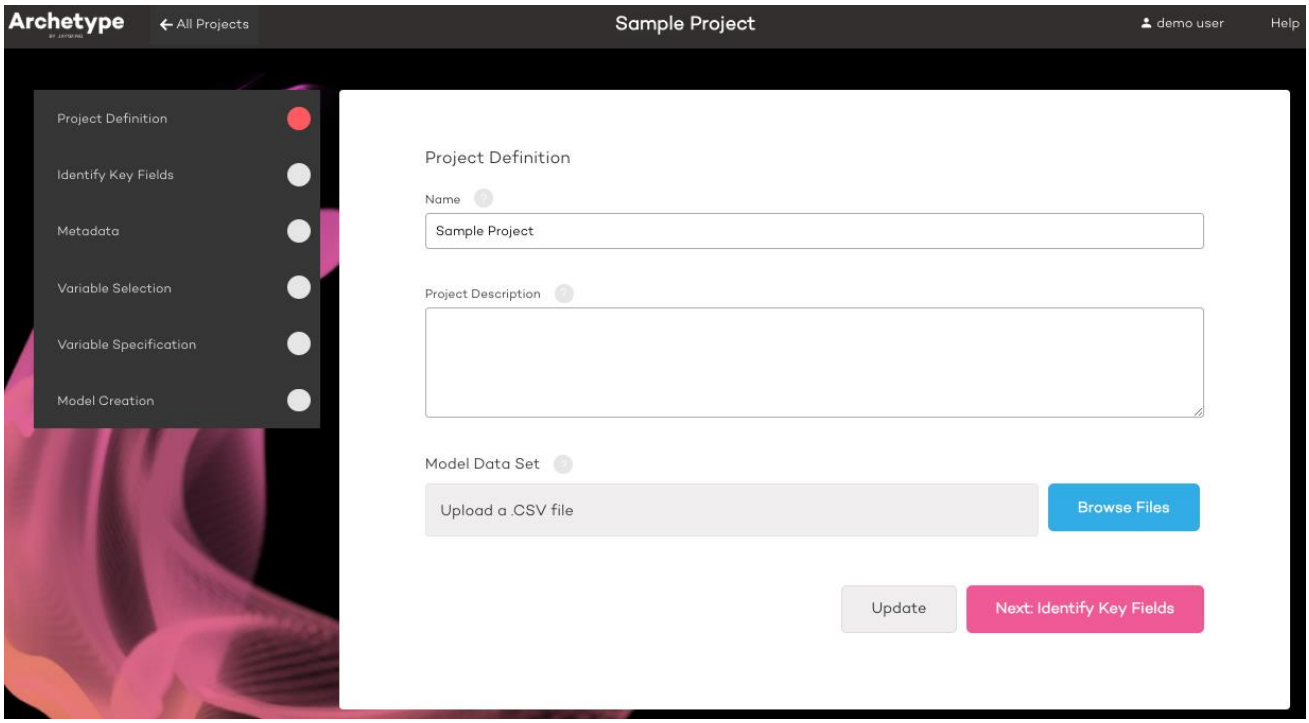
3.3 Project Definition

3.3.1 Upload Input File

At first stage of the modelling process, the model names and descriptions can be modified, and the development sample must be uploaded. The required file format is CSV. For further details, please refer to the specification table provided below.

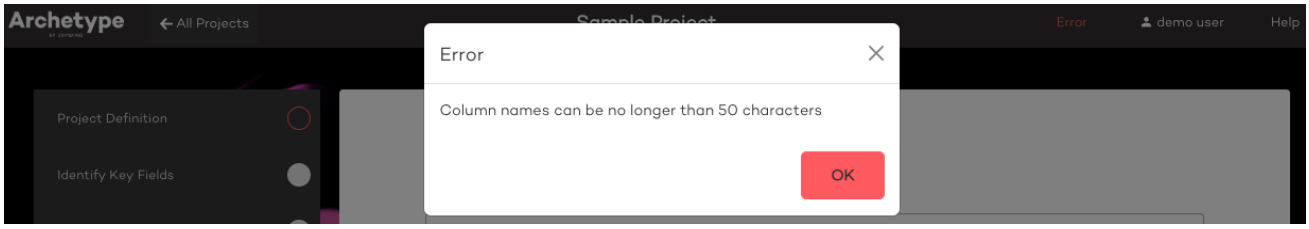
The sample file can be uploaded by clicking on the “Browse Files” button. This will open a file manager window, where the user can select the development sample file.

After the file is uploaded, Archetype performs the necessary checks to ensure all format requirements are met.



ERROR handling:

If any of the specified formatting requirements are not met, a system message similar to the one below will be displayed.

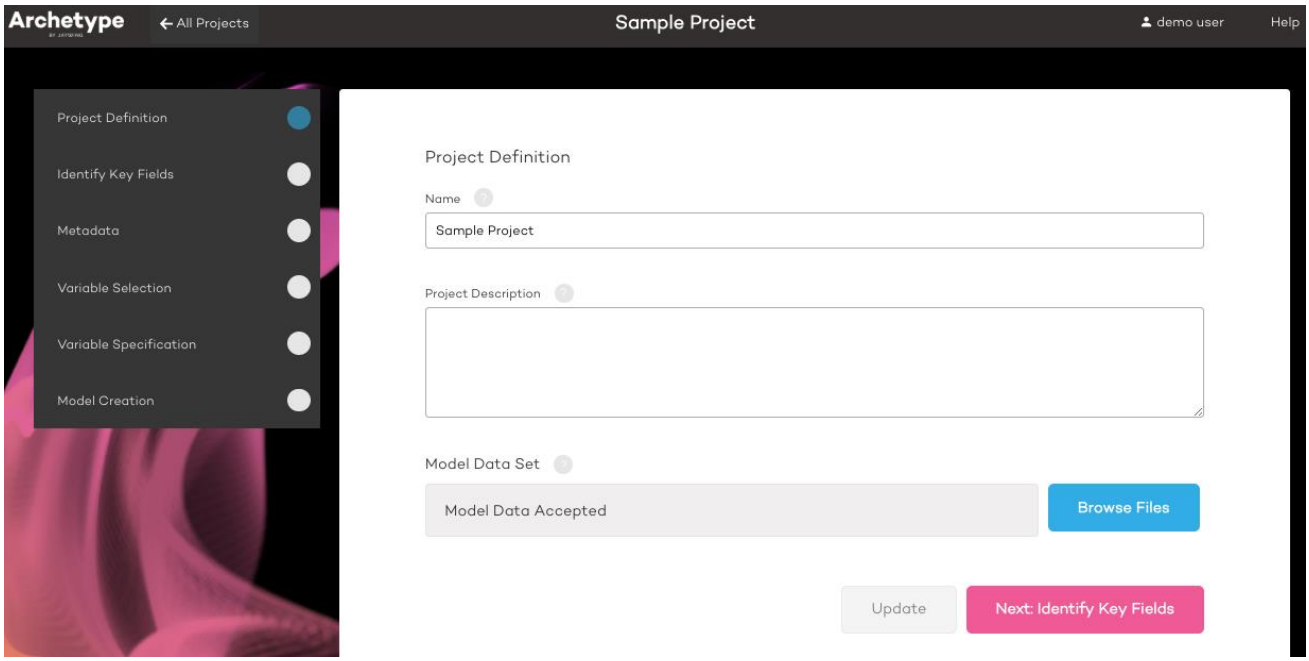


File format requirements:

Development sample file format requirements	
File extension	Csv
Delimiter	Comma
Encoding	UDF-8
Column names limit	Max 50 character
Column names	All unique. No duplication allowed.
Unique ID field	Unique value for each row/recode. No duplicates.
Outcome field	Target variable. Categorical or continuous.
Build/Validation (Test) Flag	Sample split column, with at least two categories for the identification of train and validation records. It is optional to provide test records. They won't be utilized for parameter tuning and are reserved as a true unseen sample.

3.3.2 Input file processing:

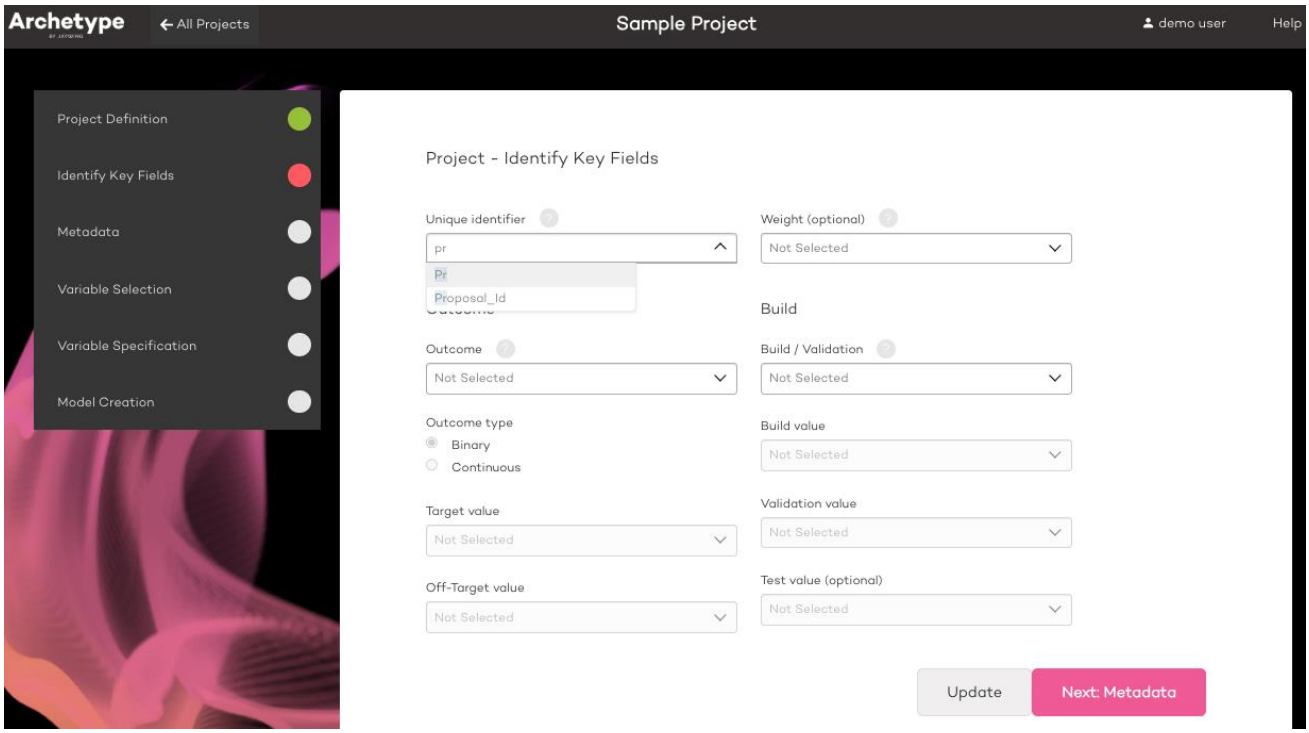
Once the input file meets all formatting requirements, Archetype will begin processing, indicated by a flashing blue light during active data processing. Users must wait for the process to finish before proceeding to the next stage. The light will turn green once the process is complete.



3.4 Identify Key Fields

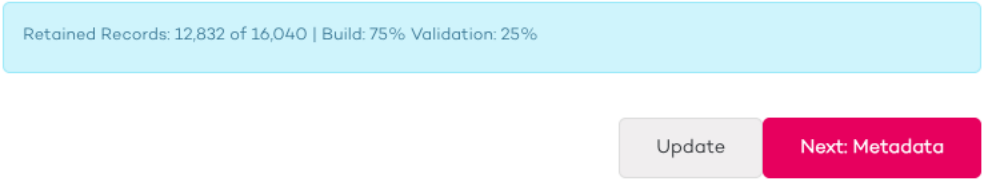
In the second stage of the modelling process, the user needs to define key data attributes, such as unique ID, outcome, and build/validation split, record weight (optional) by selecting the appropriate column header names.

The drop-down menu lists all column names and is searchable, allowing the user to type in variable names and automatically narrow down the matching columns.



Once these fields are defined, Archetype will analyse the data, and the user can use the drop-down menu (as shown below) to select the target type (and value, if it's a binary outcome) and the values representing the build and validation samples (optionally, the test sample).

After defining all key fields and clicking the “Update” button, Archetype will display the record counts retained for training and validating the model, along with their proportions. Note: In the example below, the non-retained records are from the test sample, which will not be used during model training.



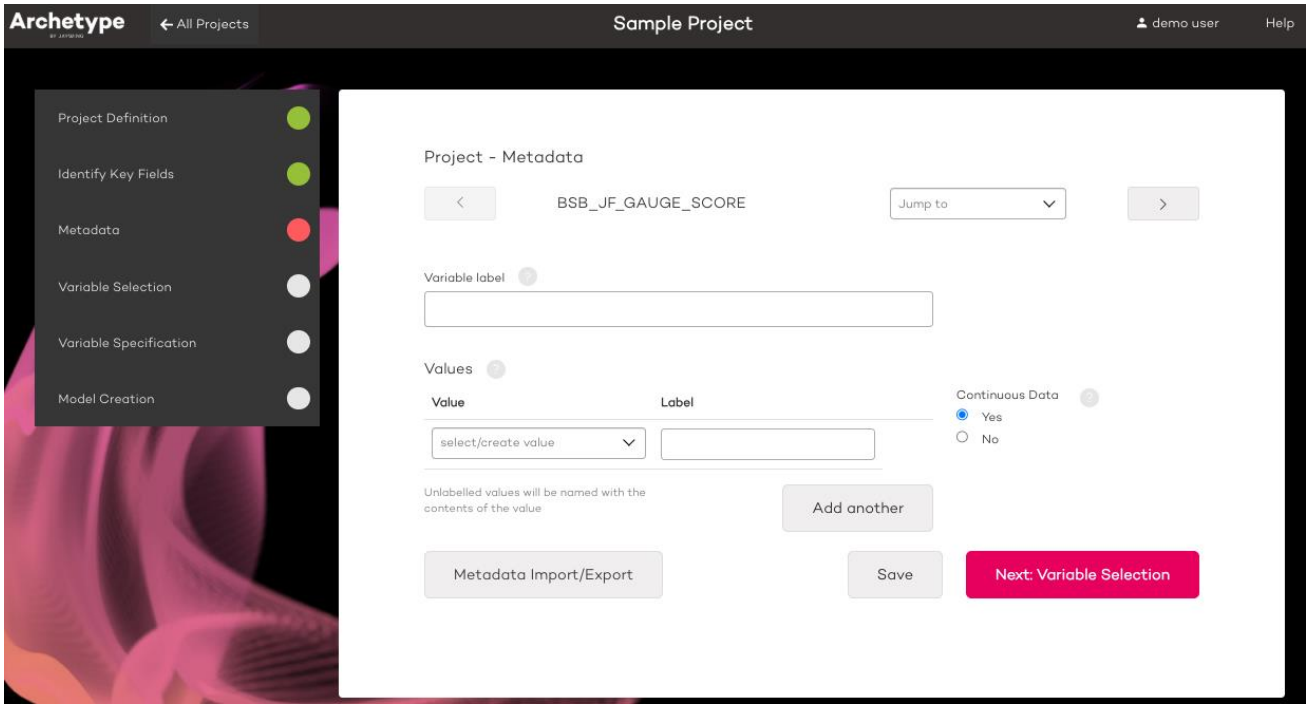
3.5 Metadata

In the metadata stage, the user can provide variable labels (descriptions), default values, and decide on whether to treat individual variables as categorical or continuous. This can either be defined manually or through (metadata) file upload (from previous projects).

If this information is not provided, Archetype will make automated decisions based on the development sample. These settings can be modified later in the project.

Note that if later modifications in the metadata lead to changes in the existing features, Archetype will reset all subsequent modelling stages to incomplete, to maintain consistency between the metadata and the model.

The user can import metadata files exported from previous projects. This file contains detailed instructions of the feature engineering applied in previous project. The file extension is “json”.

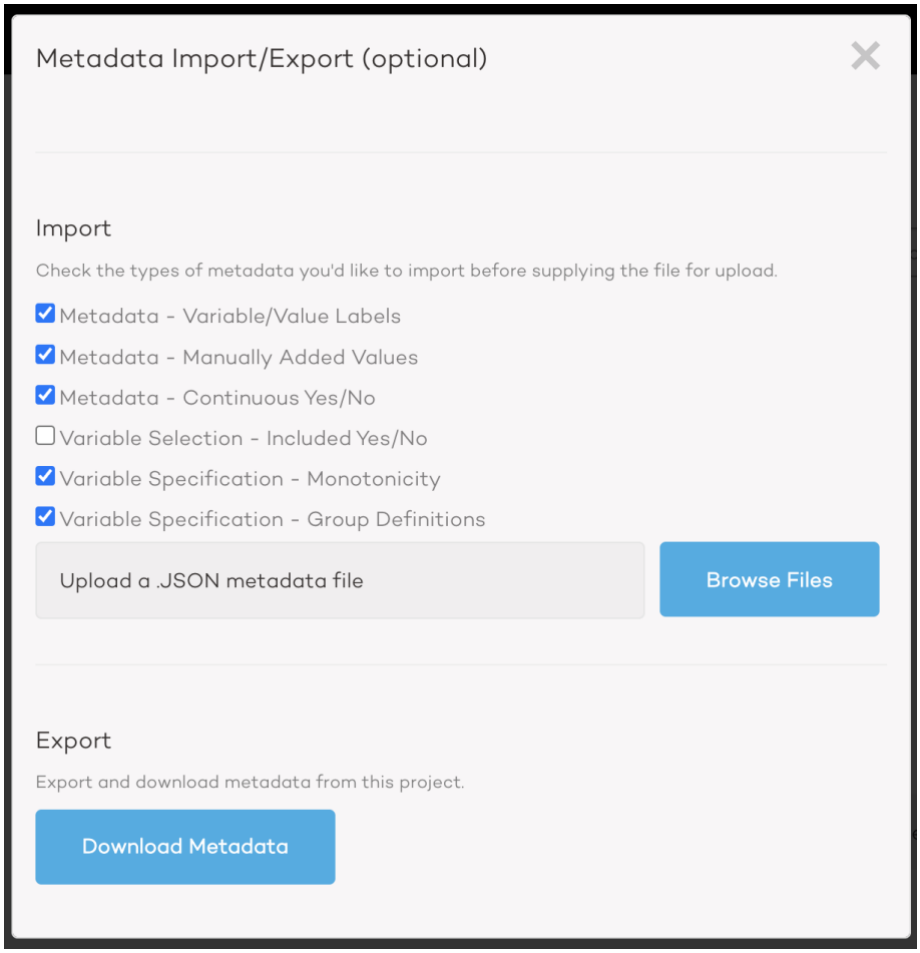


The user can define in the platform, or upload via metadata files, the following key characteristics of the data:

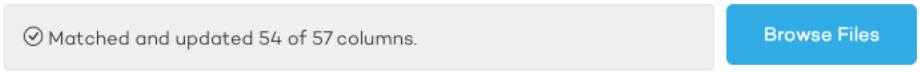
a) **Variable label**: Description of columns
b) **Variable treatment:** Categorical or continuous
c) **Value groups:** Predefined groups/bins (e.g. used for separating out default values)
d) **Value labels:** This option is useful for encoded fields. (e.g. When raw data column for occupation status is encoded as “E” in the data sample, the user can define that it appears during the model build and for documentation as well as “Employed”. The user has the flexibility to define how these codes are displayed, making it particularly useful for variables such as electoral roll bureau variables where the 1-4 integer values can be turned into meaningful descriptions).

3.5.1 Metadata Import

Metadata can be loaded (optional) by clicking the “Metadata Import/Export” button and selecting “Browse Files”. Multiple metadata files can be loaded (e.g., one for internal data and another for bureau variables). The import checkboxes can be adjusted to exclude certain aspects of the metadata from being loaded. For example, the user may choose to load only the labels (descriptions of the variables) or apply the same variable selection as of the previous model (for model refresh).



Once the metadata file import is complete, Archetype informs the user of how many columns from the development sample had matching column names, indicating a successful metadata update.

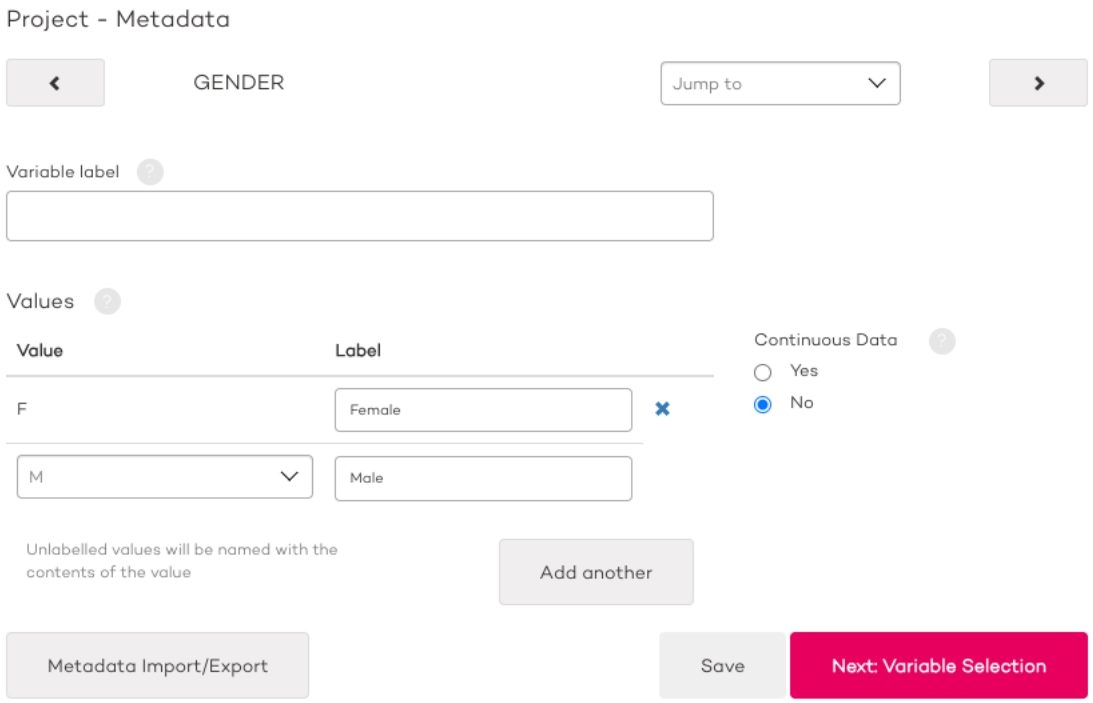


3.5.2 Metadata download

Once the feature engineering (“Variable Specification” stage) is completed, the new metadata can be downloaded by revisiting the “Metadata” modelling stage and clicking the “Metadata Import/Export” button and selecting “Download Metadata” (see image above).

3.5.3 Manual metadata definition/adjustment

The user can manually create or modify the metadata, including: adding or editing variable descriptions ("Variable label"); defining values with specific descriptions ("Label") for encoded variables (e.g., F → Female); treatment of data (continuous/categorical). See description of key action buttons below.



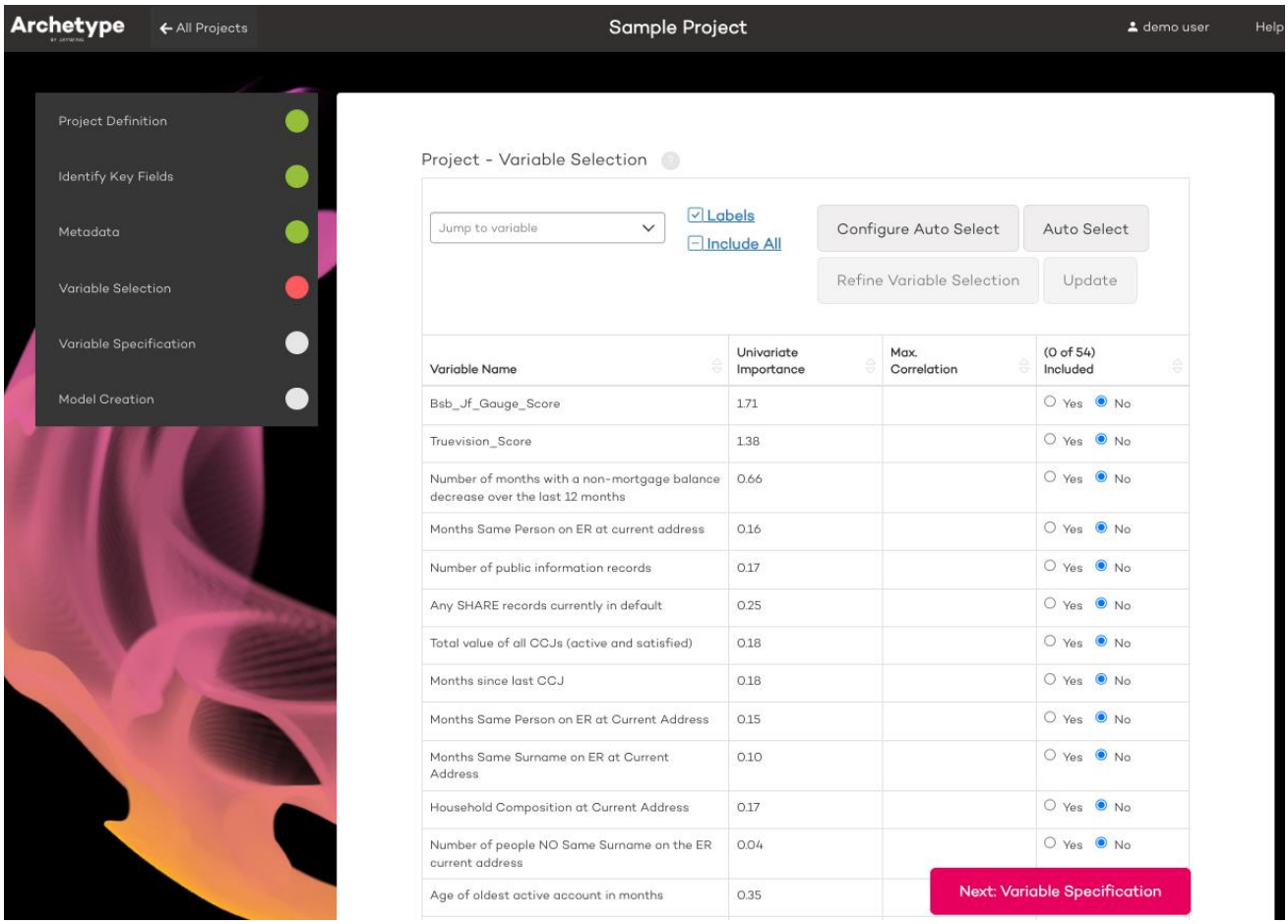
Action Buttons	Description
Jump to:	Searchable drop-down menu for columns.
Navigation bars (“<” “>”)	Navigation to the next or previous column.
Variable label	Variable description this will appear in the model output.
Value	Drop-down menu to select input value and define label.
Add another	Add further values and labels.
Continuous Data	Specify whether the data should be treated as continuous or categorical. Note: If no metadata is uploaded, Archetype automatically infers this based on the sample.
Save	Save down column definitions. Required to save changes before navigating to another column.

3.6 Variable Selection

The goal of this modelling stage is to remove unwanted variables from the feature list and reduce model complexity by using fewer features. This can be done manually or automatically by the software, based on user-defined rules (e.g. maximum correlation, minimum variable importance, or a set limit on the number of variables).

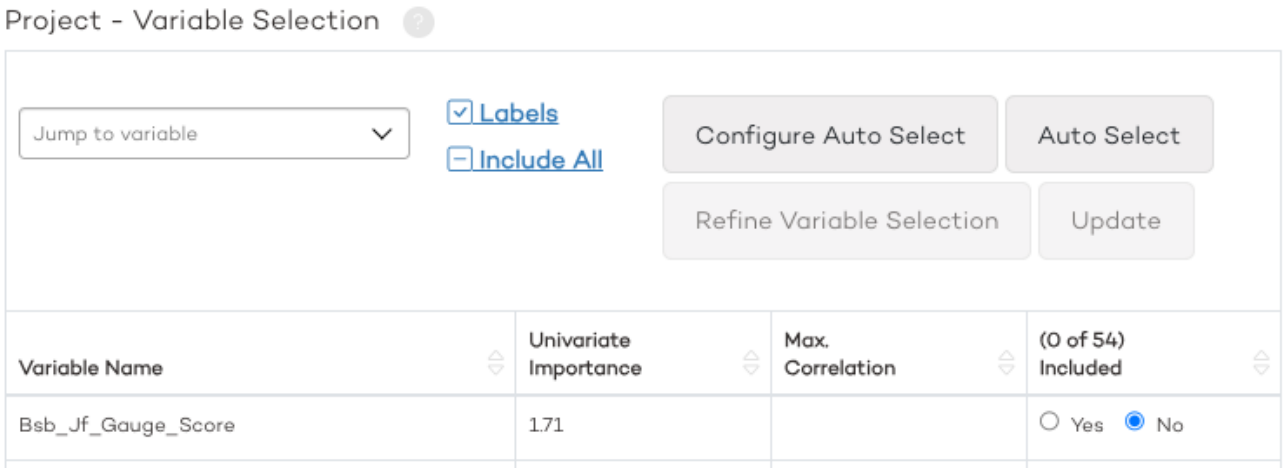
Archetype supports variable selection by providing key metrics for each variable, such as univariate importance (information value), maximum correlation and feature importance (once a model has been trained).

The order of the variable list can be adjusted by sorting each column in ascending or descending order, allowing the modeller to achieve an efficient and thorough review of the variables.



3.6.1 Variable Selection Functions

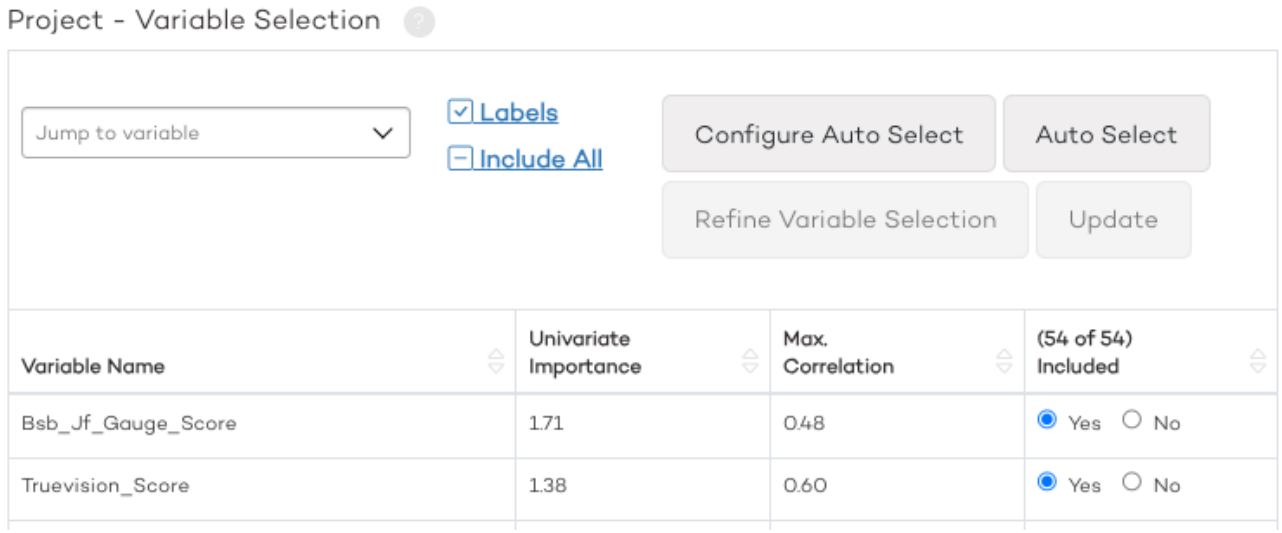
The buttons in the "Project Variable Selection" panel offer useful options that assist the model developer in efficiently selecting model variables



Function	Description
Jump to variable	User can select a variable from drop down menu to jump to the statistics of the relevant variable, therefore limiting scrolling
Labels	Whether to use variable descriptions (labels) or the raw input column names (headers of the uploaded sample)
Include All	Selects all features
Configure Auto Select	Define auto variable selection rules
Auto Select	Run automatic variable selection based on the configured rule sets
Refine Variable Selection	Change automated variable selection rules (once auto select completed)
Update	Refresh correlation matrix based on current selection
Included Yes/No	Indicates whether the variable is selected for model training. Set automatically when “Auto Select” completed but can be also adjusted manually by the user.

3.6.2 Manual Selection

The “Included: column shows whether a feature has been selected for modelling. Users can choose to select or deselect individual features, or use the “Include All” button to select or deselect all features at once.

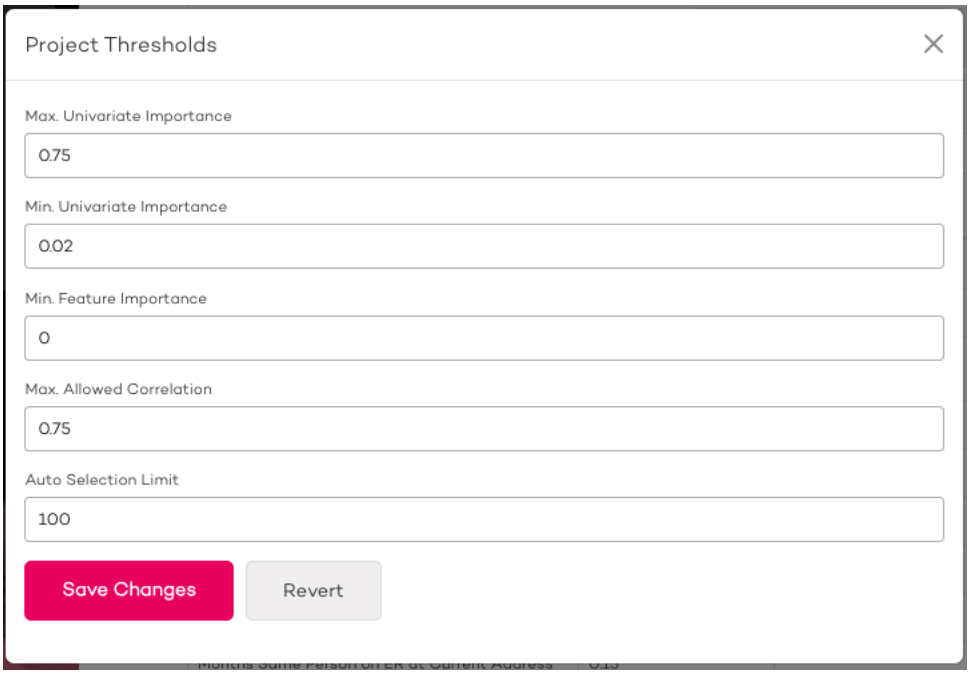


3.6.3 Correlation matrices

Top correlating variables are displayed when hovering over max correlation values, helping to identify highly correlated variable pairs.

3.6.4 Configure Auto Select

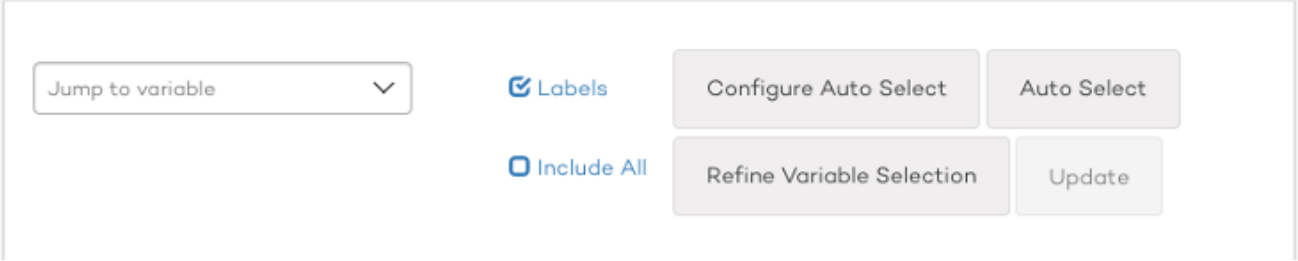
Auto selection rules can be set/modified by clicking on the “Configure Auto Select” button. The settings are prepopulated with default values, which can be modified by typing in new values into the fields.



Function	Description
Max. Univariate Importance	Variables exceeding this univariate importance threshold will be excluded from selection. A common use case is preventing data leakage.
Min. Univariate Importance	Variables below this value are excluded.
Min. Feature Importance	Minimum threshold based on relative feature importance of the trained model.
Max. Feature Importance	Maximum threshold.
Max. Allowed Correlation	Correlation threshold, the less important feature will be deselected by the software.
Auto Selection Limit	The maximum number of selected variables.

3.6.5 Auto Select

"Auto Select" trains an initial model and applies filtering logics defined in the Auto Select configuration.



3.6.6 Feature Importance

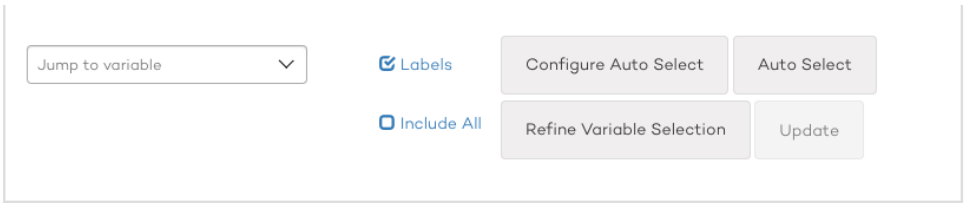
Once the initial model has been built, a new column with Feature Importance will be displayed. It shows relative importance compared to most important variables (100).

The table can be sorted ascending or descending based on each column, further supporting manual reviews (e.g. features with highest correlations or lowest univariate importance)

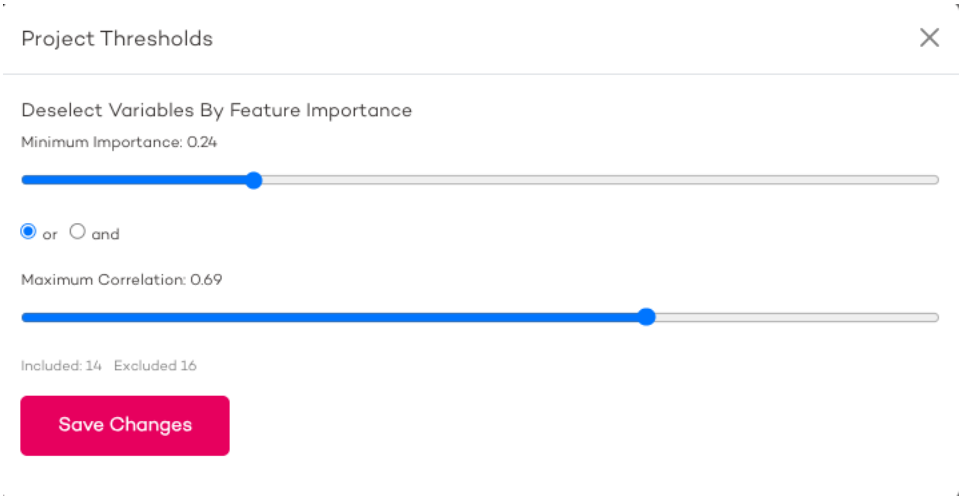
Variable Name	Feature Importance	Univariate Importance	Max. Correlation	(30 of 54) Included
Bsb_Jf_Gauge_Score	0.29	1.71	0.24	Yes No
Truevision_Score	0.29	1.38	0.30	Yes No
Number of months with a non-mortgage balance decrease over the last 12 months	0.28	0.66	0.29	Yes No
Months Same Person on ER at current address	0.16	0.16		Yes No
Number of public information records	0.13	0.17		Yes No
Any SHARE records currently in default	0.03	0.25		Yes No
Total value of all CCJs (active and satisfied)	0.17	0.18	0.25	Yes No

3.6.7 Refine Variable Selection

The “Refine Variable Selection” button can be used after running “Auto Select” to apply further filters on the feature selection.

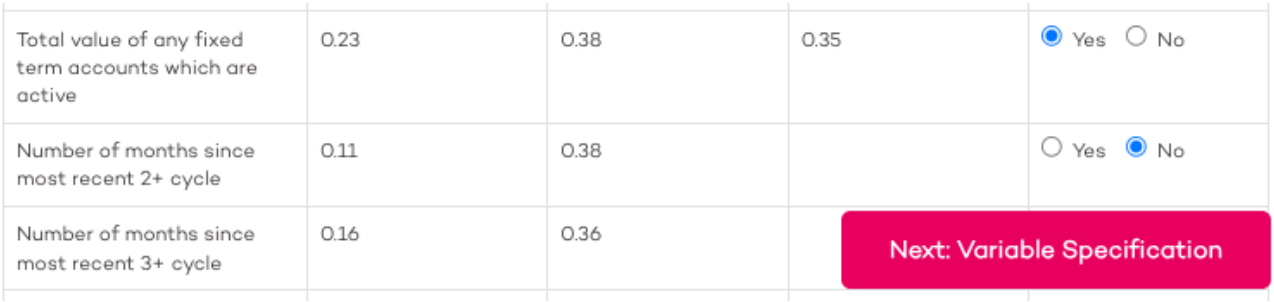


Two parameters, minimum feature importance and maximum correlation, can be added with an or/and relationship. Archetype immediately calculates the included and excluded variable counts when the sliders are moved.



3.6.8 Next: Variable Specification

Once the initial variable selection is ready, click on “Next: Variable Specification”. Note that the user will be able to return to this stage and refine variable selection if needed. A blue flashing light in the progress bar on the left hand side next to the “Variable Specification” will indicate that the data are processed for the next modelling stage.



3.7 Variable Specification

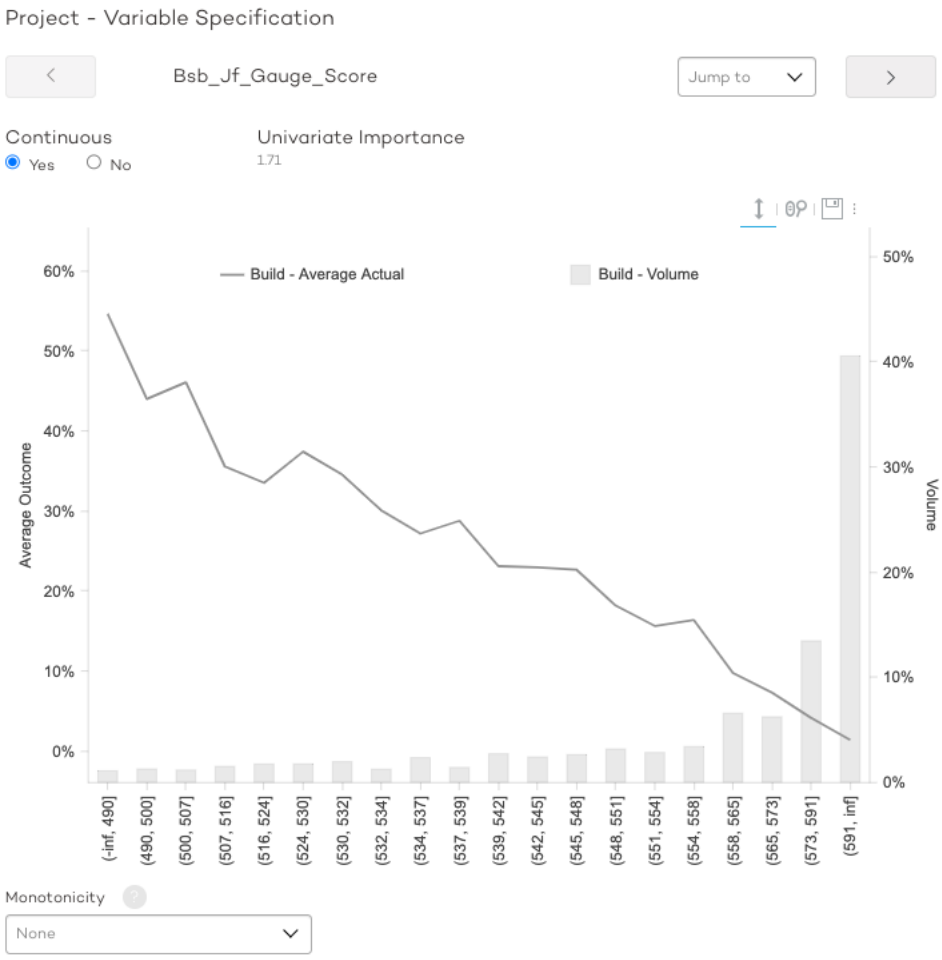
3.7.1 Summary

Variable specification or feature engineering is an important part of robust model development. This modelling stage supports further variable review as well the creation of user defined features/groups (coarse classes).

The software provides partial dependency plots (PDP) for each selected variables, which help to understand data distributions, and the relation between variable values/groups and the outcome value (“Average Outcome”) of value groups.

The user can apply monotonicity constrains, ranking constrains of categorical variables as well change the variable type bet ween continuous and categorical.The chart below is an example of the PDP chart provided. The bars represent the weighted record proportion for each bin (group) whilst the markers line shows the average target outcome.

Note: For continuous variables these bins are created only for illustration purposes.

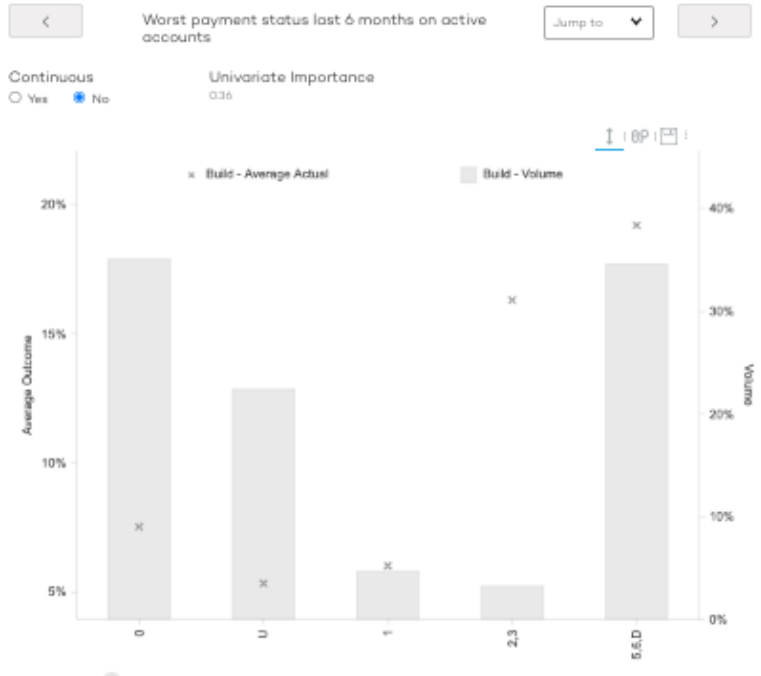


3.7.2 Value Groups / Classes

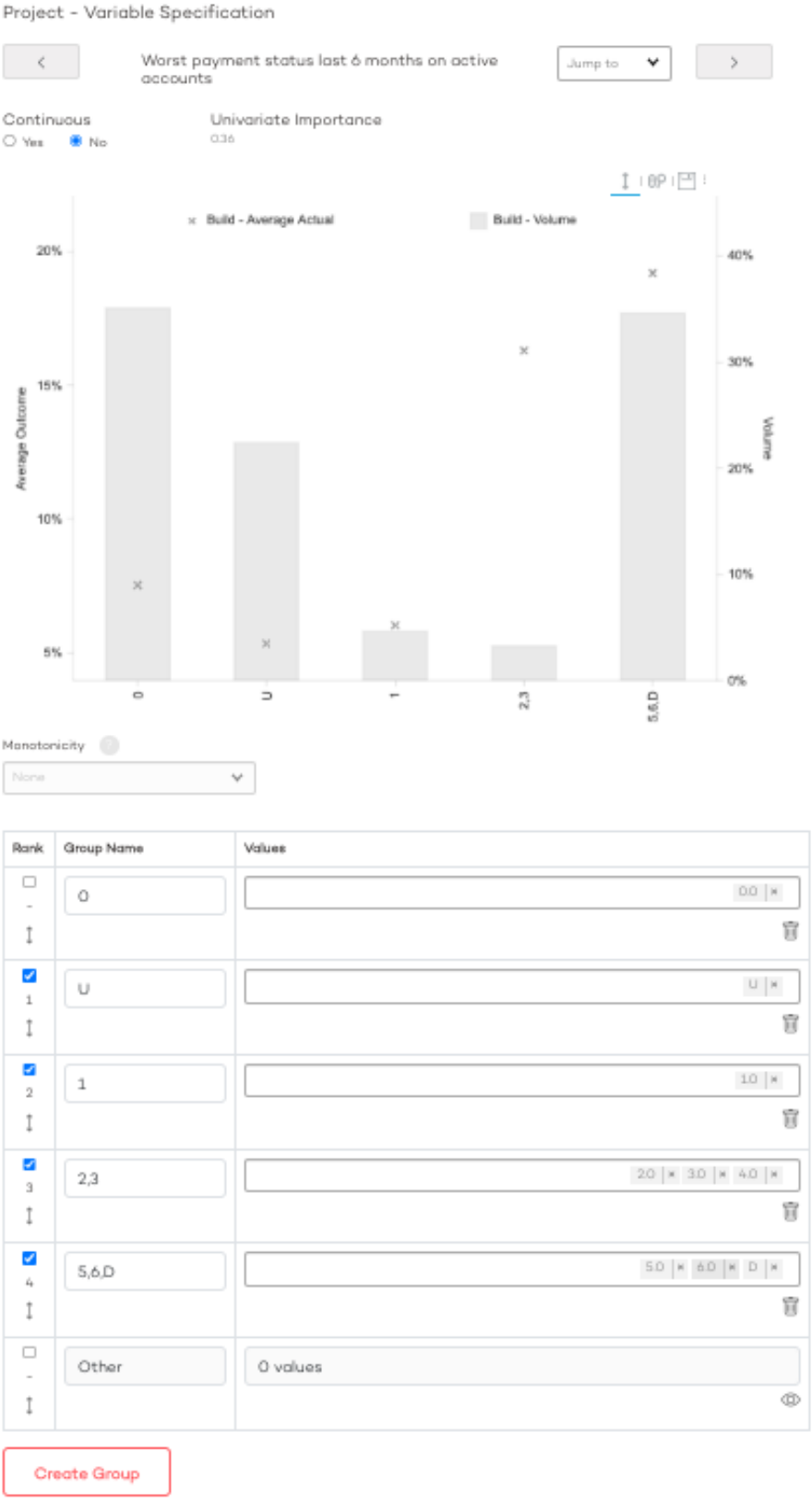
The user can separate out special values, such as default values of bureau data, as well can create coarse classes by binning different values together to achieve more intuitive trends (example below).

Chart before applied coarse classing:

Chart post coarse classing



Archetype's no-code solution allows user to create these new groups through some simple clicks, in the group creation panel provided below each PDP chart.



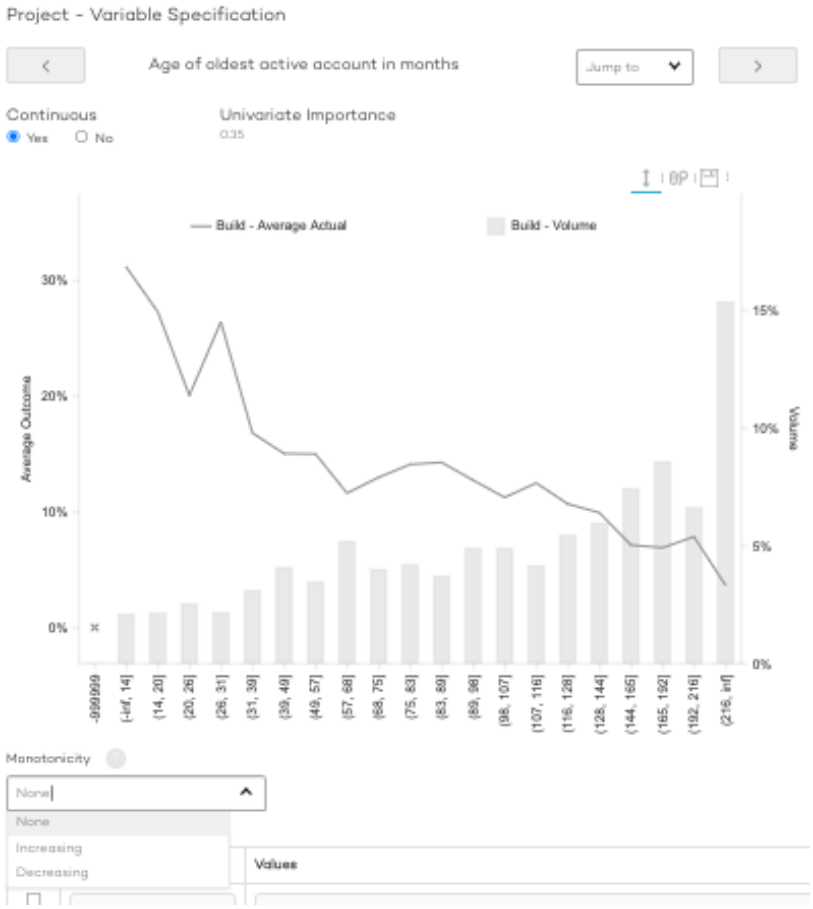
3.7.3 Save

Make sure after any changes made in the groups that you save your changes by clicking on the “Save” button.

3.7.4 Monotonicity

The relation between the variable values and the model prediction can be enforced, by selecting the desired direction (“Increasing” or “Decreasing”) in the “Monotonicity” drop-down menu.

When selected Archetype enforces the direction of the marginal impact of the variable. Note: this will be applied on the continuous part of the variable. For categorical groups/values ranking order can be defined.



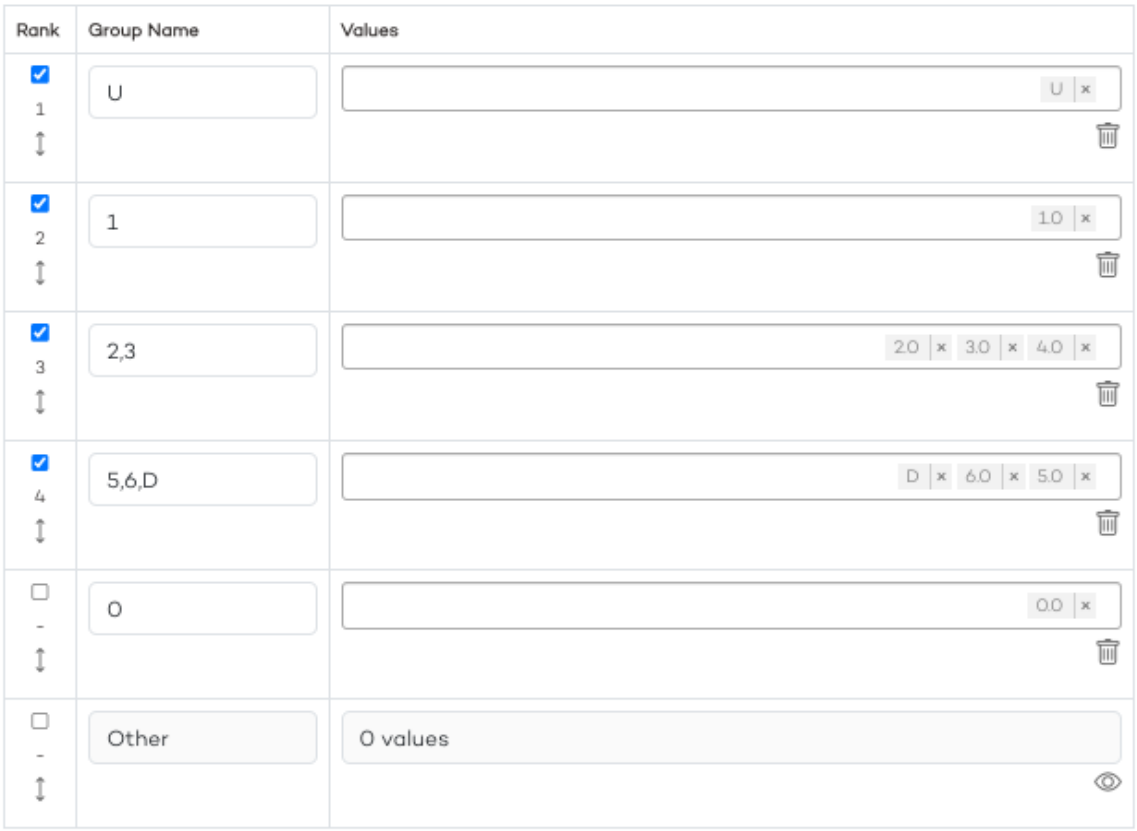
None	No monotonicity applied
Increasing	Increased variable value results in non-decreasing model prediction
Decreasing	Increase in the variable value results in non-increasing model prediction

3.7.5 Rank

Monotonicity constraints can also be applied to categorical variables by ticking the rank box on the left side of the variable group.

Rank orders can be adjusted by clicking and dragging the group to the desired position in the list.

The ranks represent increasing monotonicity (i.e., a higher rank value corresponds to a higher predicted value).



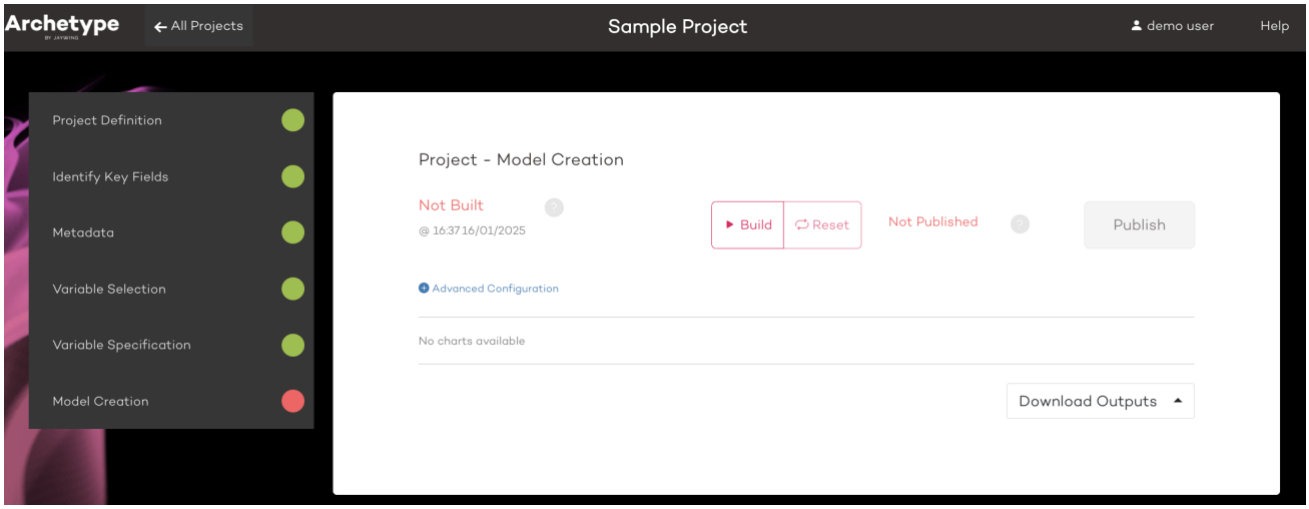
3.7.6 Next: Model Creation

Once thefeature engineering and variable review in the ‘Variable Specification stage is completed click on “Next: Model Creation” to progress to the next modelling stage.

3.8 Model Creation

The “Model Creation” stage is the final phase of the model development process, dedicated on training the model.

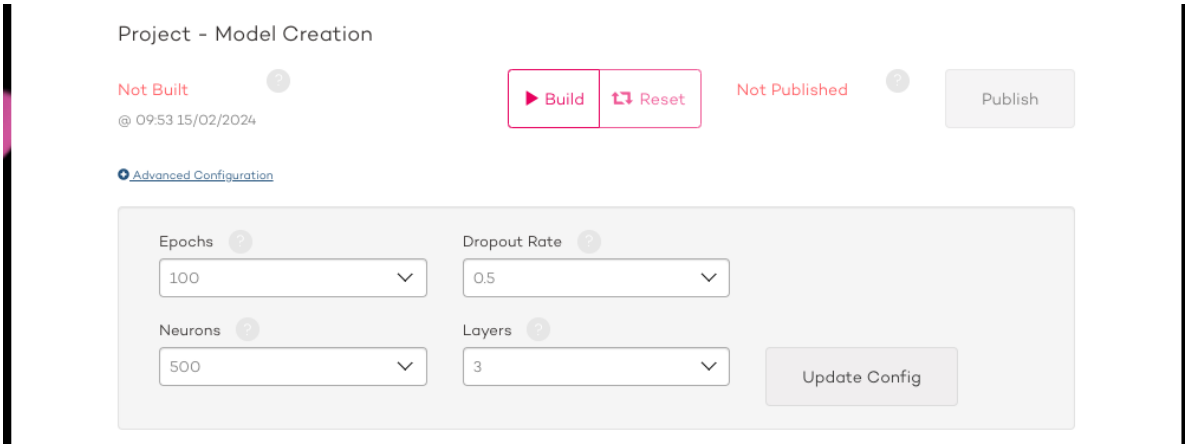
At this stage, all the traffic lights should be green except for “Model Creation”. Key functions are listed in the table below.



Build / Pause Build	Build: Start model training. If model already trained it allows to train it for further iterations. Pause: Pause model build, available during model training is running.
Reset	Reset the trained model, removing all learned parameters.
Publish	Lock the model and generate a published version, which will be added to the “Model” list.
Advanced Configuration	Setting model parameters (e.g. number of layers, neurons, etc.)
Download Outputs	Provides a drop-down menu for downloading key project information, including: a) Partial dependency plot charts b) Column information (IV, correlation, feature importance) c) Feature creation code

3.8.1 Advanced Configuration

Clicking the “Advanced Configuration” button opens a drop-down panel (shown below). The model architecture can be defined by selecting options from the drop-down menu. After choosing the parameters, click the “Update Config” button to apply the changes.

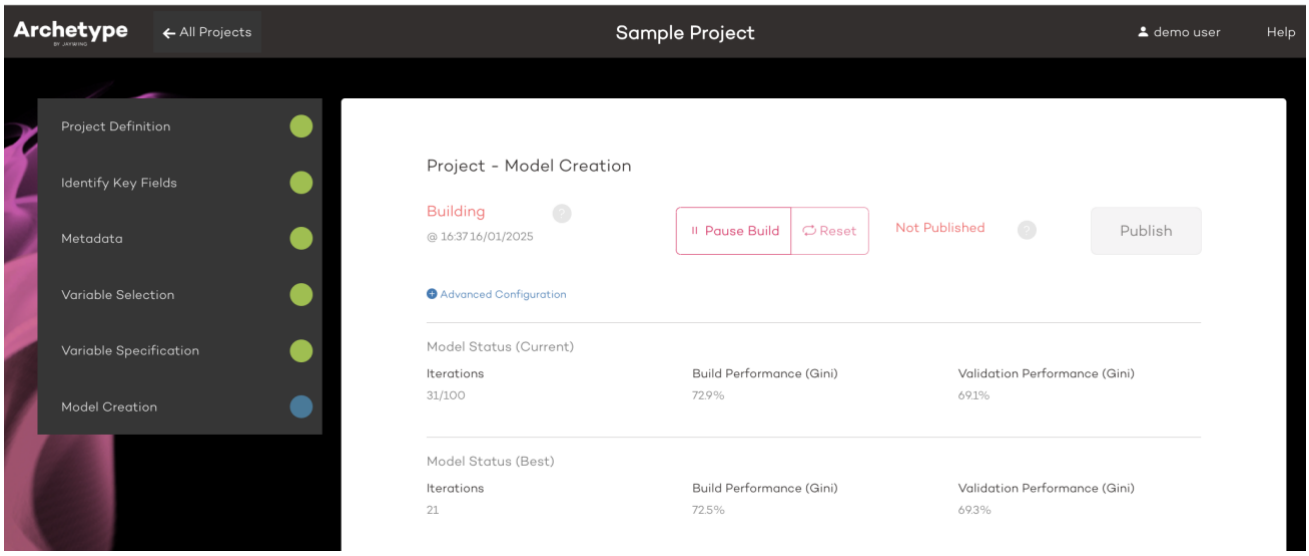


Epochs	Number of epochs the model should be trained on.
Dropout Rate	Reset the trained model, removing all learned parameters.
Neurons	Number of neurons per layers. *Note: Each layer will have the same number of neurons. For example, if 500 neurons are selected with three layers, the model will be built with 500 neurons in the first, second, and third layers.*
Layers	Dept of the architecture. Number of hidden layers.

3.8.2 Build

Once “Build” button is clicked Archetype prepares the data for model training applying normalisation, finding optimal learning rate and learning scheduler.

During model build the “Build” button changes to “Pause Build” which allows to pause the training process. During training the last iteration’s performance (“Current”) and the best model performance is shown.



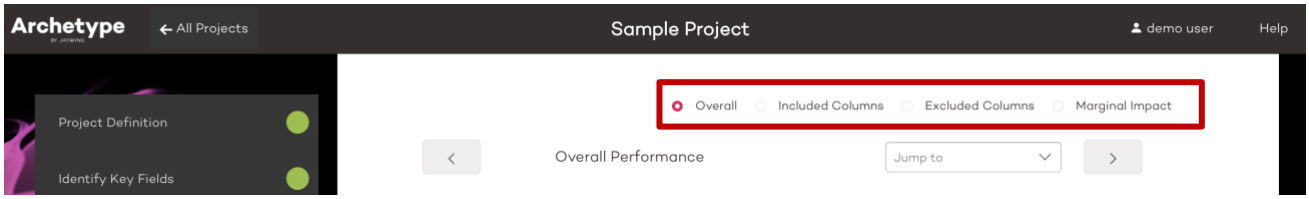
Once the build process has run (or been paused), Archetype will return the best model based on the performance metrics from the validation sample (e.g. Gini for binary outcome and R2 for continuous outcome). It also calculates performance on the test sample when provided.

Model Status (Best)

Iterations	Build Performance (Gini)	Validation Performance (Gini)	Test Performance (Gini)
22	72.3%	69.5%	68.6%

3.9 Model Outputs

Once the model training is finished, Archetype creates supporting statistics and charts to help model review and documentation.



Model outputs are represented in four tabs:

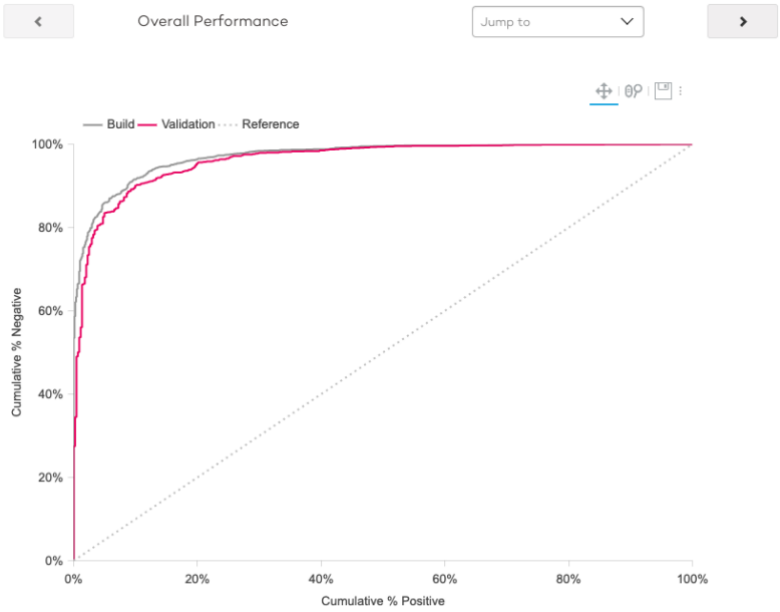
Overall	Visual representation of overall performance such as: - ROC curve - feature importance - marginal variable importance.
Included Columns	Partial dependency plots (PDP) based on actuals and predictions for each model feature (on build and validation sample).
Excluded Columns	PDP for excluded columns.
Marginal Impact	ICE (Individual Conditional Expectation) charts displaying how a model's predictions change for a specific instance as one feature varies, while keeping all other features constant.

3.9.1 Overall Performance

Overall performance charts are visual representation how well the model discriminate and predict.

Receiver operating Characteristic curve (ROC) curve (binary outcome):

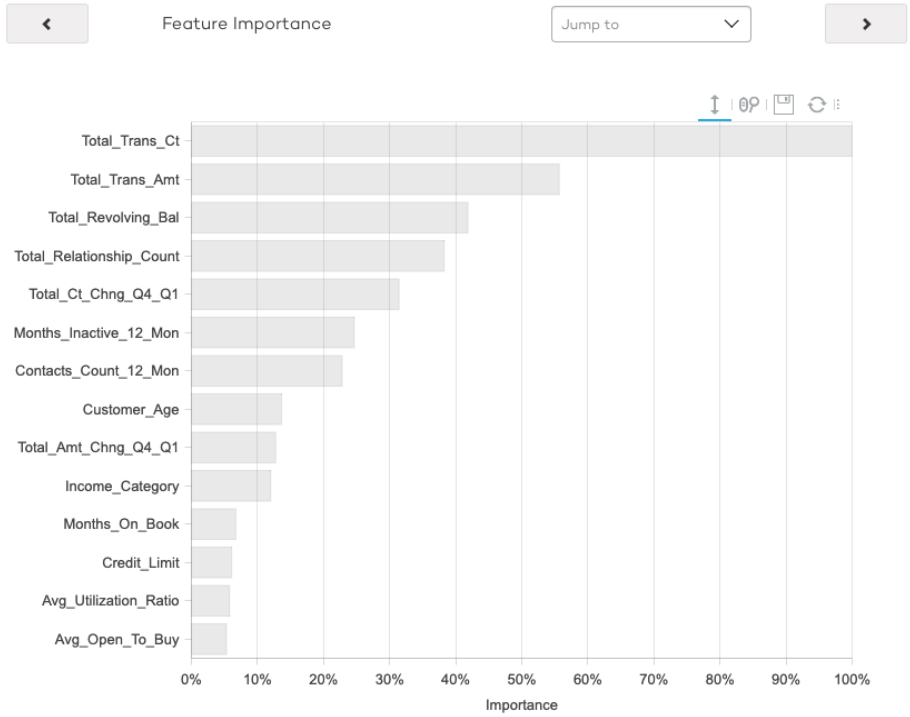
Represents the separation power between target and non-target based on Build and Validation samples.



Prediction vs Actual Scatter Plot (continuous outcome)

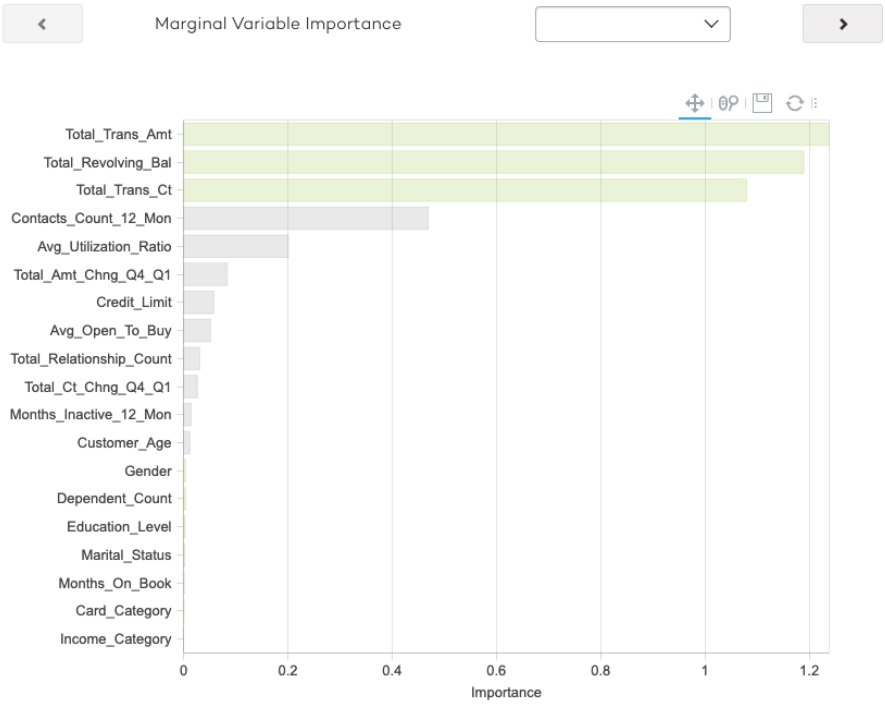
3.9.2 Feature Importance

A visual representation of the relative importance of the top predictors relative to the most important variable.



3.9.3 Marginal Variable Importance

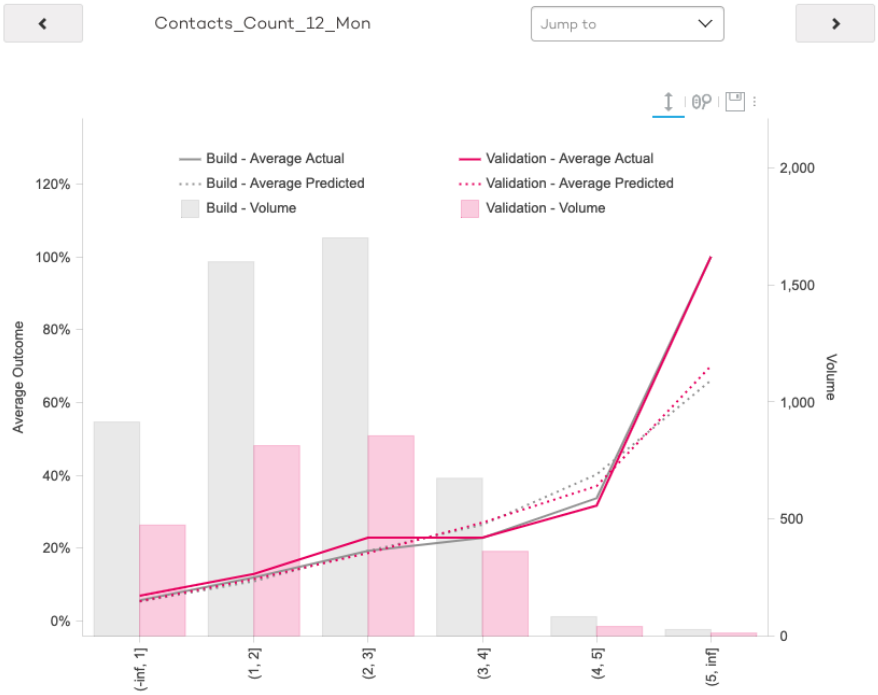
Marginal importance illustrates the individual impact of each variable. Variables included in the model are displayed in grey, while excluded variables are shown in green.



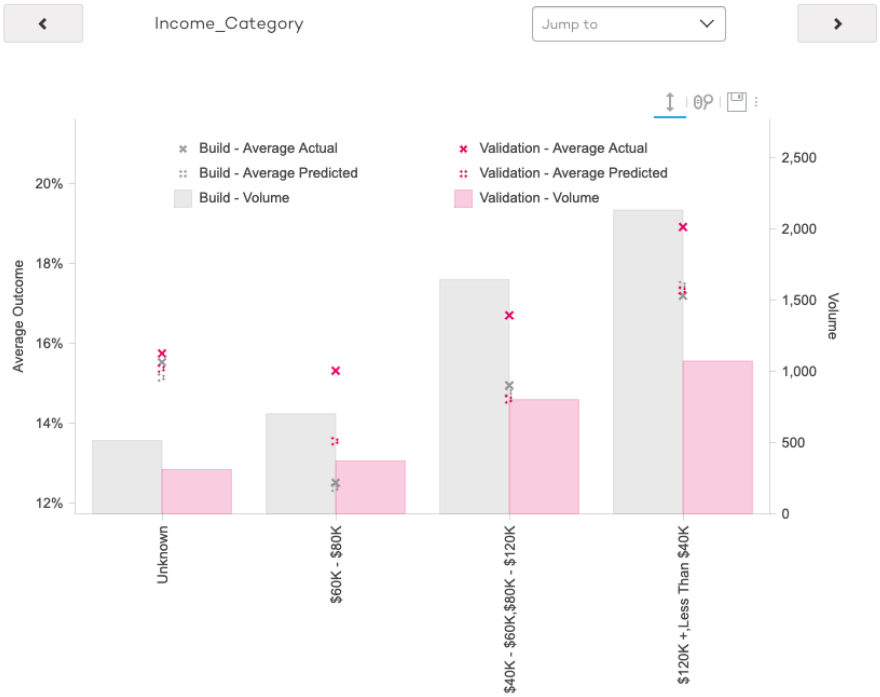
3.9.4 Included Columns

Partial dependency plots are generated for each column included in the model, displaying both actual values and predictions. Values from the training sample are shown in grey, while those from the validation sample are presented in pink.

Example chart of continuous column:



Example chart of categorical column:



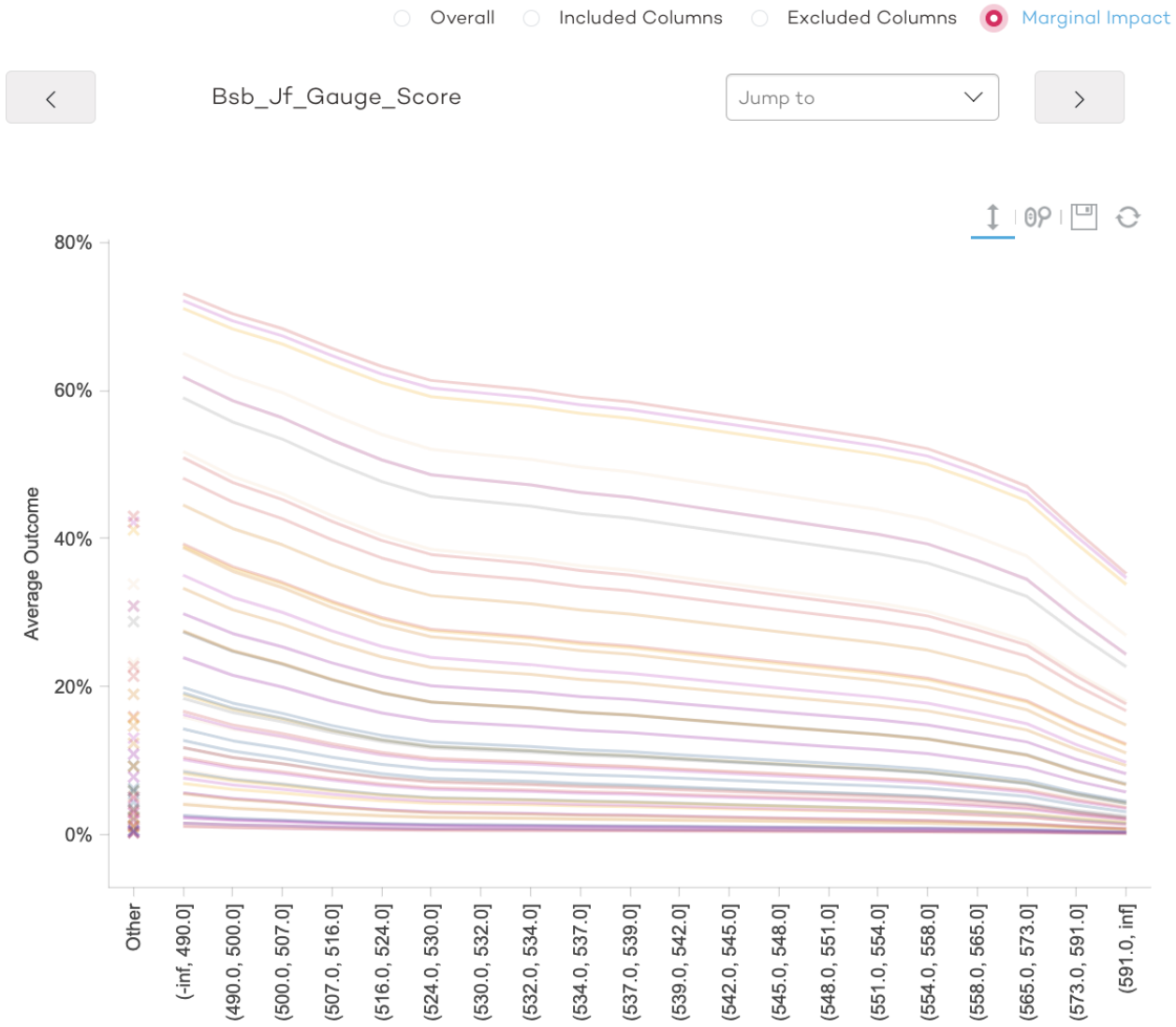
3.9.5 Excluded Columns

This tab displays the PDP charts for the excluded columns, using the same representation as for the included columns.

3.9.6 Marginal Importance

ICE (Individual Conditional Expectation) charts displaying how a model's predictions change for a specific instance as one feature varies, while keeping all other features constant.

The following example should be interpreted as indicating that the model produces lower predictions for each individual account as the Gauge Score increases, while keeping all other factors constant.

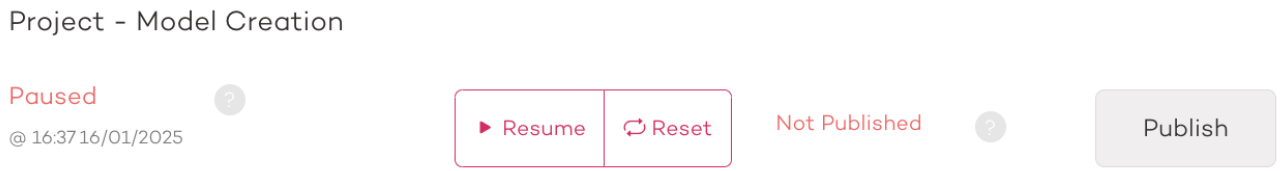


3.10 Publish Model

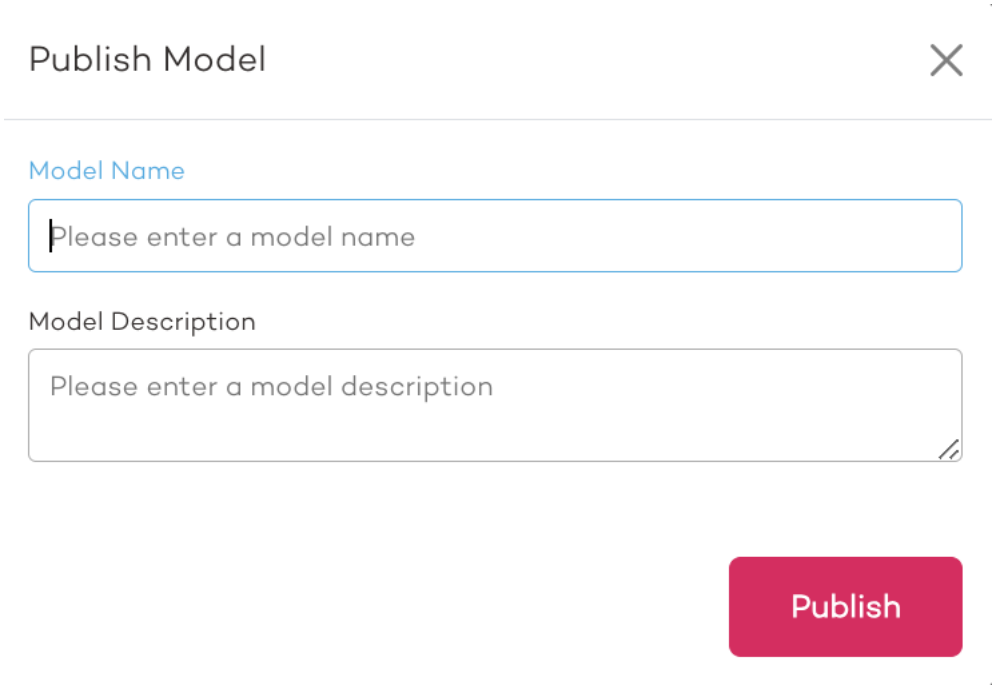
The final step is to publish the model, which locks down the trained model.

The model can be published by clicking on the grey “Publish” button.

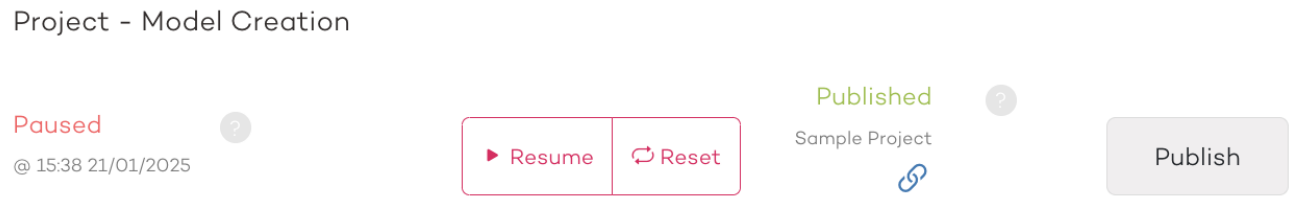
Note “Not Published” red sign indicates the current status of the model.



When the "Publish" button is clicked, a pop-up window will prompt you to enter the "Model Name" and a description for the model to be published.

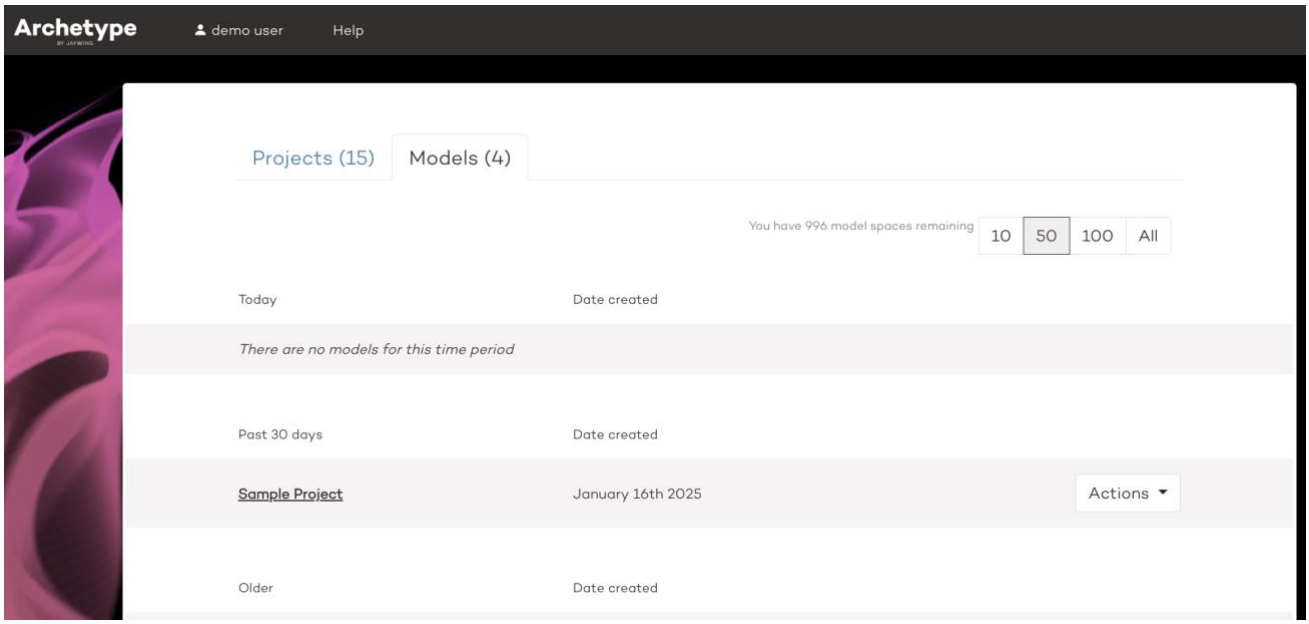


Once the model is published, a green "Published" text indicates its status, with the published model name displayed underneath. The blue link symbol next to the model name is an active link, when clicked, it directs you to the published model site.



4 Models

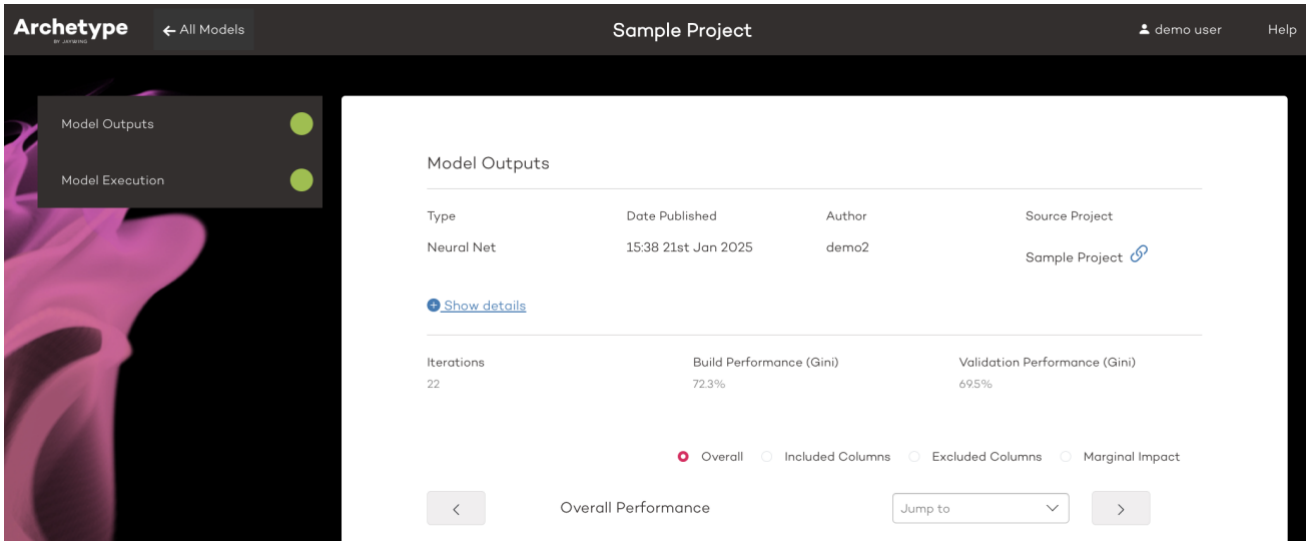
Once a project is published, it is added to the "Models" list, which is the second tab on the Main page (next to the "Projects" tab). This list displays the published models in a similar fashion and with the same functionality as the "Projects" tab.



The published models contain all information required for model deployment, as well as documentation and inferencing (i.e. calculating predictions based on new samples).

Each model will have two main pages:

- **"Model Outputs:** including key information as well deployment codes
- **"Model Execution":** to run model prediction on new samples, as well extracting impact factors.

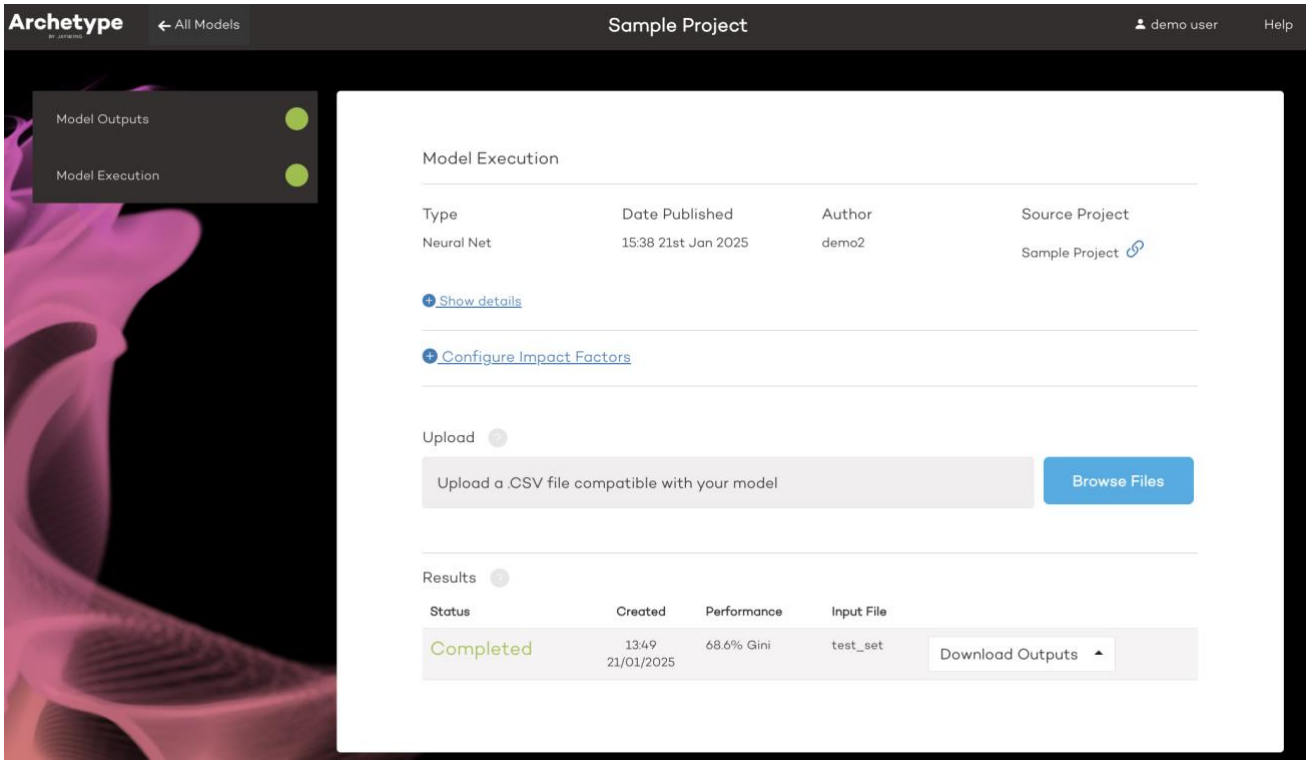


4.1 Model Outputs

The “Model Outputs” tab includes details of the model's architecture, publication date, author information, and a hyperlink to the original project. The charts and statistics created during model development are available for display and download.

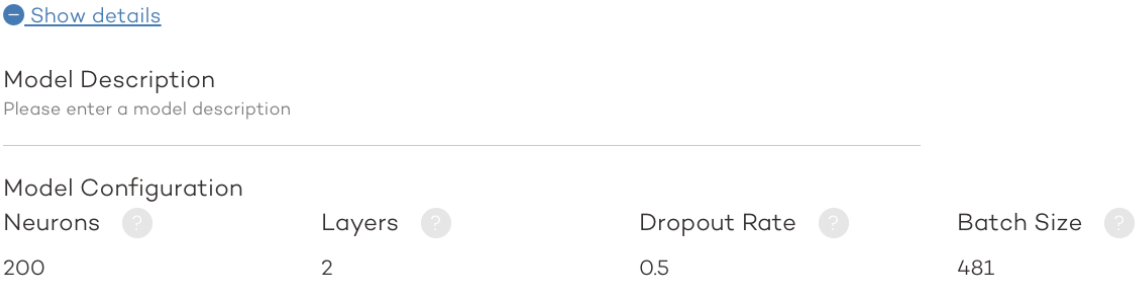
4.2 Model Execution

The “Model Execution” tab allows the user to create predictions based on new data samples, as well calculating impact factors (i.e. variables which had the most impact on individual predictions).



4.2.1 Show Details

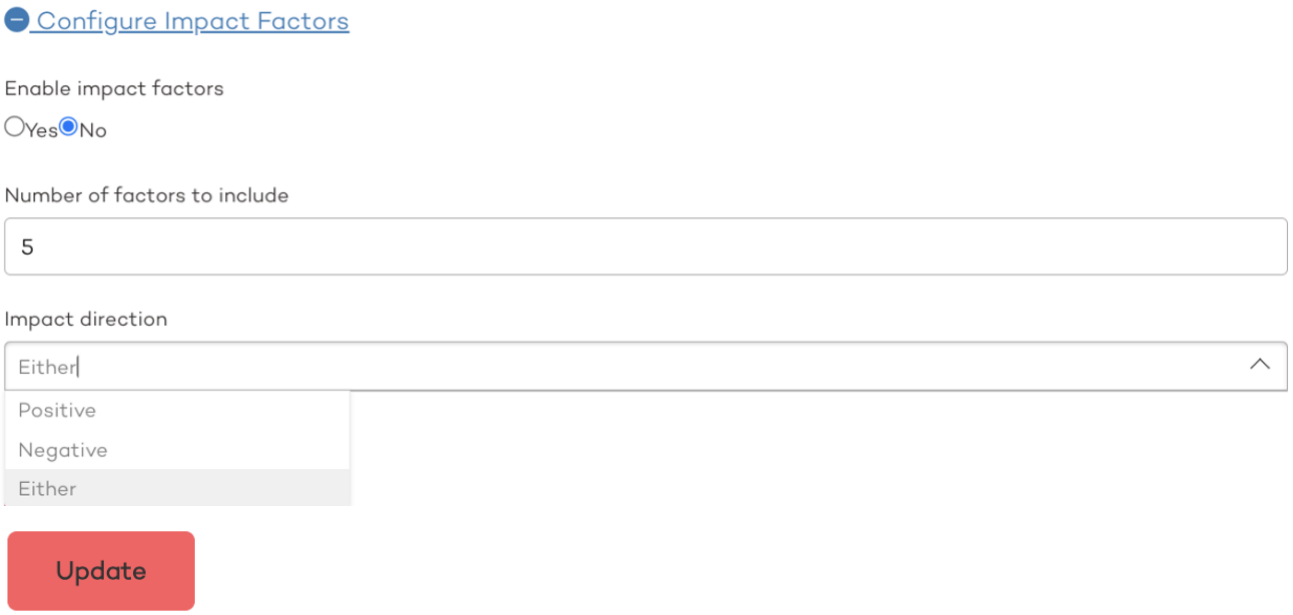
“Show Details” is an exampandablepanel, displaying model architecture.



4.2.2 Configure Impact Factors

In the “Configure Impact Factors” panel the user can enable and disable the calculation of impact factors, as well as updating the configuration.

If the “Enable impact factors” switch is set to “Yes”, Archetype will calculate and output the impact factors. By default, it is set to “No”. The user also can define how many factors should be listed in the report (“Number of factors to include”).

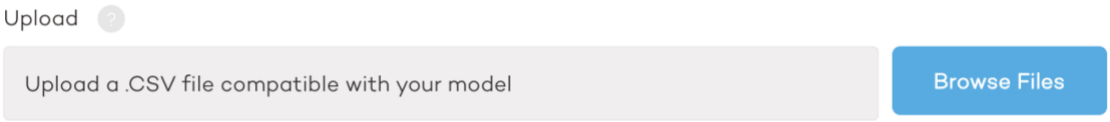


Configurations:

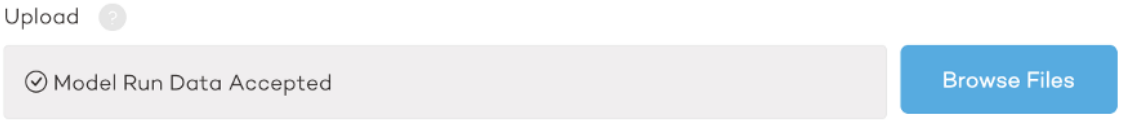
Enable impact factors	default this is setto (“No”), meaning impact factors are not calculated.
Number of factors to include	The number of most influential variables displayed in the output. Note: All factors are still computed.
Impact direction	**“Positive”:** Displays only factors/variables that contributed to an increase in the prediction (e.g. in a fraud scorecard, those that raised the probability of fraud). **“Negative”:** Displays only factors/variables that lowered the prediction. **“Either”:** Shows the top impact factors, regardless of whether they increased or decreased the prediction.
Update	Make sure to click on “Update” to apply the changes.

4.2.3 Upload Files

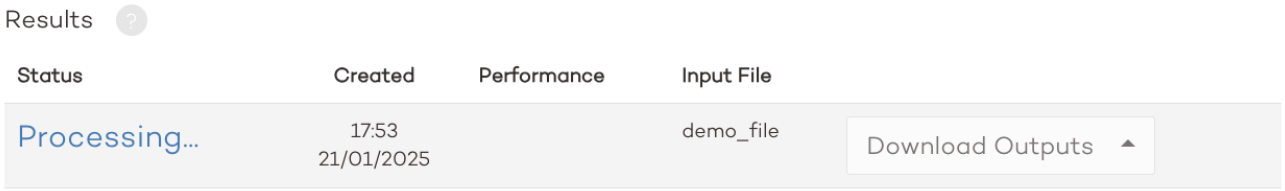
New samples can be uploaded via clicking on the “Browse Files” button and selecting the file from the file explorer.



The file format must match the model development sample, including the unique ID header. Archetype verifies compliance and ensures all feature columns are present. If valid, it displays the message below and begins generating predictions, and new line will be added to the “Results” tab with the file name.

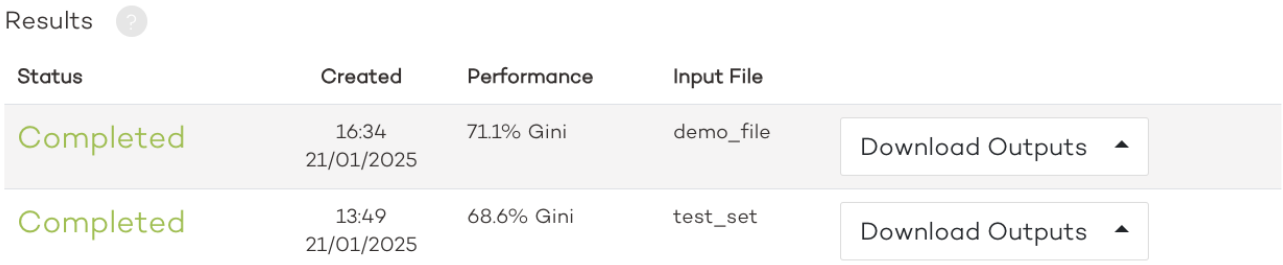


Once file is accepted the status of the report will be displayed as shown below (“Processing”).

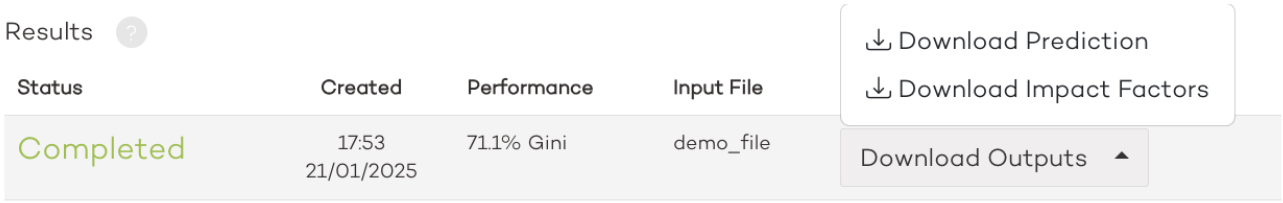


4.2.4 Prediction Reports

Once completed, the Status changes to “Completed.” If the outcome column with the target value were provided overall performance metrics will be displayed too.



Predictions can be downloaded from “Download Outputs” drop-down menu. If impact factors are enabled, a second report for impact factors will also appear.



Output Examples

Below are examples of predictions and impact factor reports.

Example of Prediction csv

	A	B	C
1	Clientnum	Attrition_Flag	Predictions
2	768805383	Existing Customer	0.000294077
3	818770008	Existing Customer	4.92E-05
4	713982108	Existing Customer	7.23E-06
5	769911858	Existing Customer	0.000575759
6	709106358	Existing Customer	6.52E-06
7	713061558	Existing Customer	0.002003205
8	810347208	Existing Customer	0.002183803

Example of Impact Factors csv

	A	B	C	D	E	F	G	H	I
1	Clientnum	original_predictio	feature_1	delta_1	prediction_1	original_value_1	feature_2	delta_2	prediction_2
2	768805383	0.03%	Total_Ct_Chng_Q4_Q1	-0.24%	0.27%	1.625	Avg_Utilization_Ratio	-0.17%	0.20%
3	818770008	0.00%	Total_Ct_Chng_Q4_Q1	-0.18%	0.18%	3.714	Avg_Utilization_Ratio	-0.04%	0.05%