IYKRA
Data Fellowship Program

Predictive Analysis with Dataiku

by Rizki Fajar Nugroho

(date of delivery)

Trainer Profile

Rizki Fajar Nugroho
Data Scientist at SaaS Company
LinkedIn

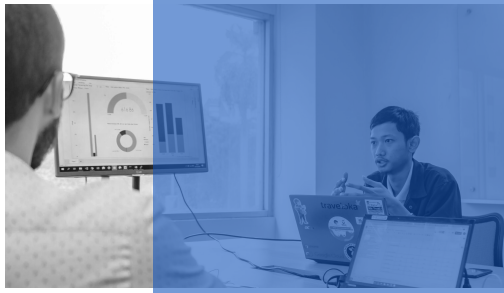
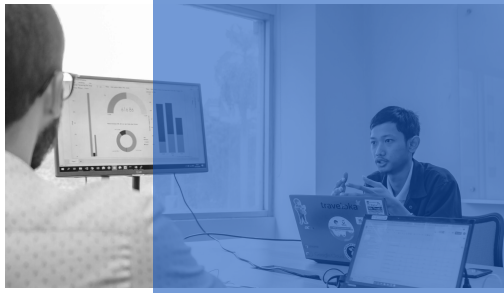


Table of Content

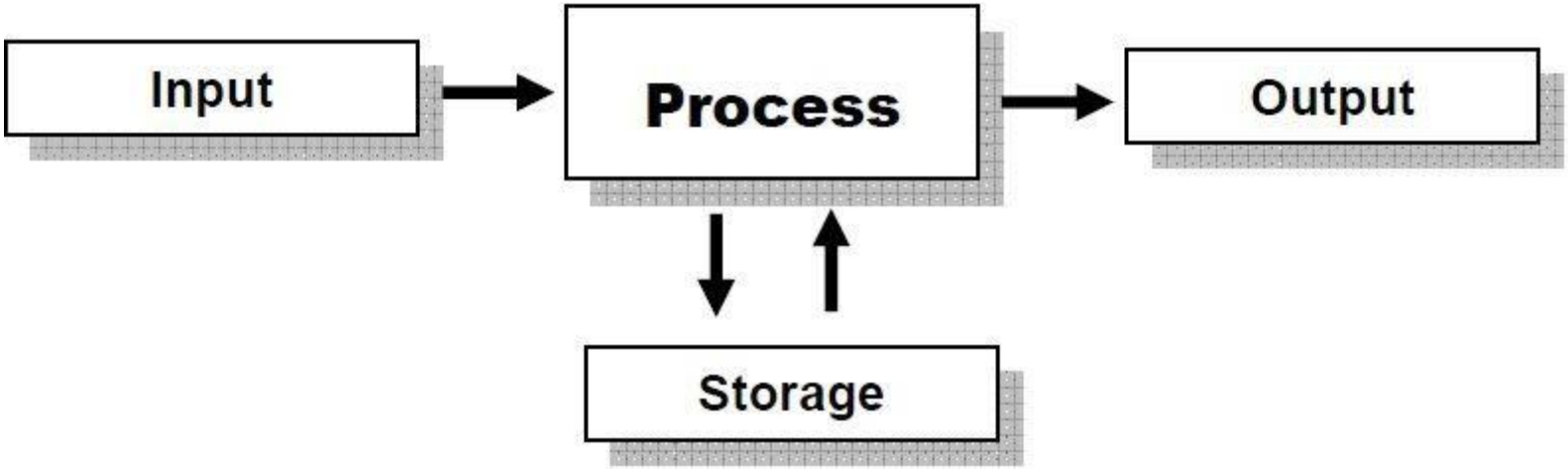


Data Fellowship

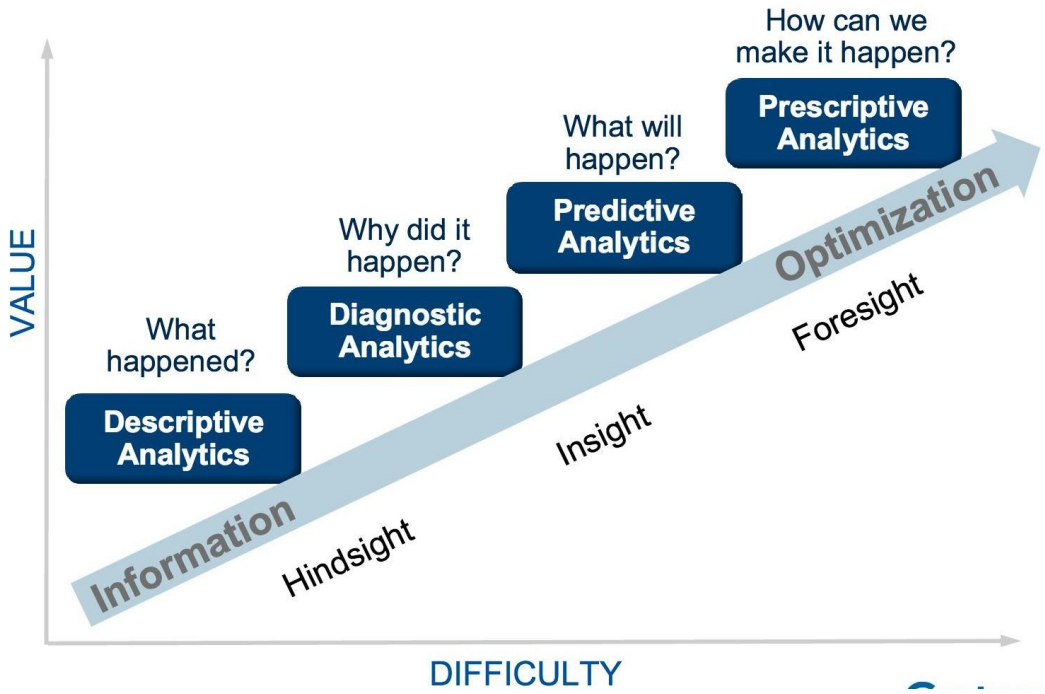
Intro to Predictive Analytics

What is Data Analytics

is **science and techniques** of analyzing **raw data** in order to **make conclusions** about that information



Type of Analytics



Predictive Analytics Examples

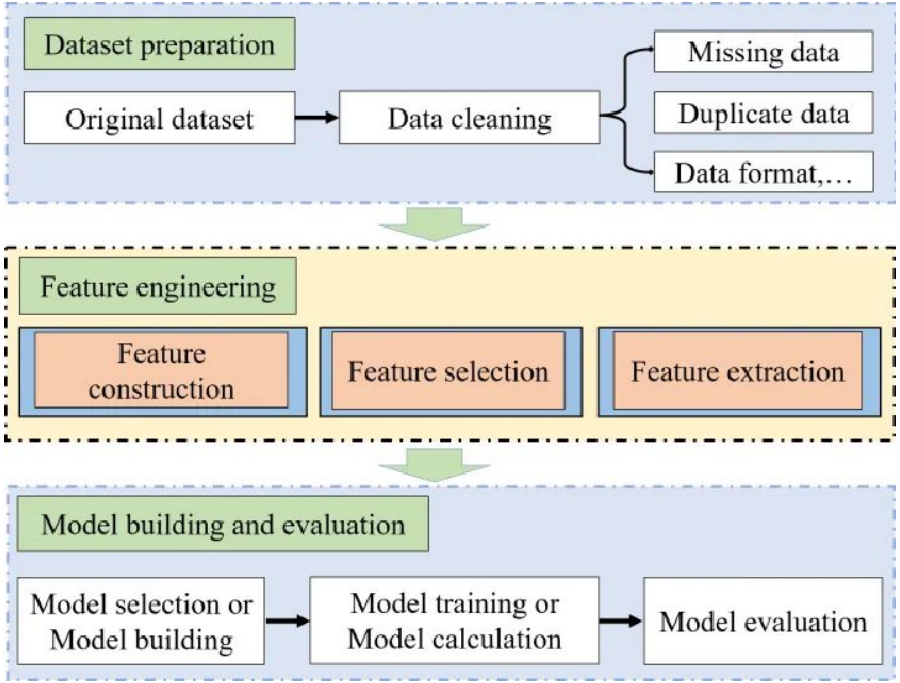
In the banking sector, predictive analytics is paramount for informed decision making across various facets:

- Customer Behaviour Analytics: By analyzing past transactional data and customer interactions, banks can predict future financial behaviors and tailor personalized offerings and services to individual customers, enhancing customer satisfaction and retention.
- Credit Risk Assessment: Predictive analytics models assess creditworthiness by analyzing historical financial data, enabling banks to make informed lending decisions and mitigate the risk of default.

Predictive Analytics Examples

- Fraud Detection: By monitoring transactional patterns and detecting anomalies in real-time, predictive analytics helps banks combat fraudulent activities, safeguarding customer assets and maintaining trust.
- Market Trends Forecasting: Predictive models analyze market trends, interest rates, and economic indicators to anticipate shifts in the financial landscape, enabling banks to adapt strategies and capitalize on emerging opportunities.

Machine Learning Breakdown Steps



Data Fellowship

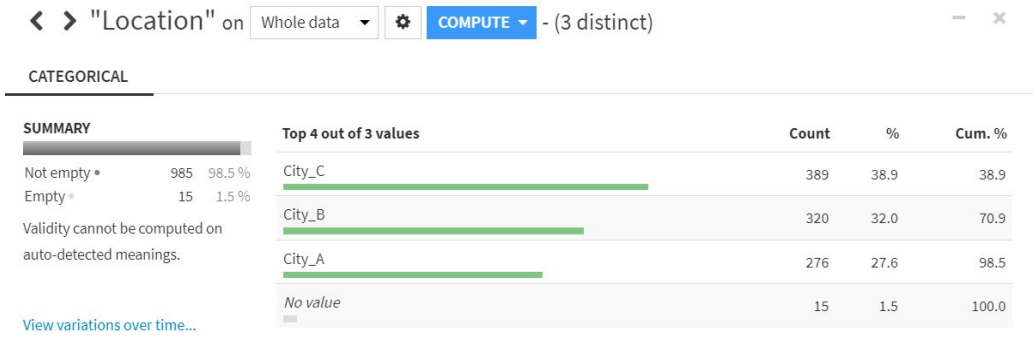
ML Workflow in Dataiku - Data Preparation

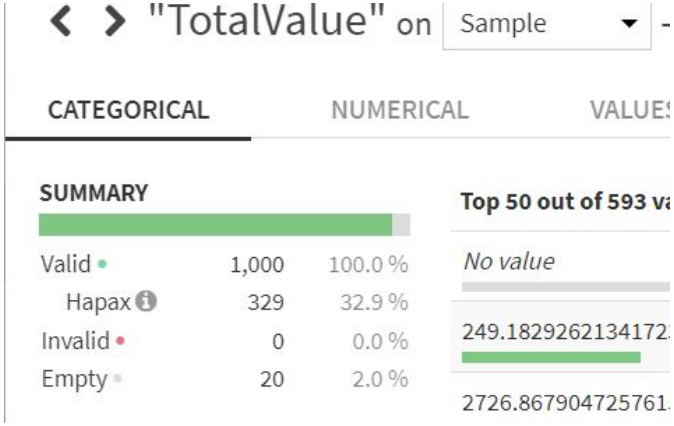
Importance of Data Preparation

- Effective data preparation is essential for ensuring the quality, consistency, and reliability of data used in predictive analytics.
- It involves cleaning, transforming, and structuring raw data to make it suitable for analysis and modeling.
- Proper data preparation can significantly impact the accuracy and performance of predictive models
- Dataiku provides a wide range of visual recipes for data preparation, making it easy for users to perform common data transformation tasks without writing code and it could be done through **visual analyses** on Dataiku

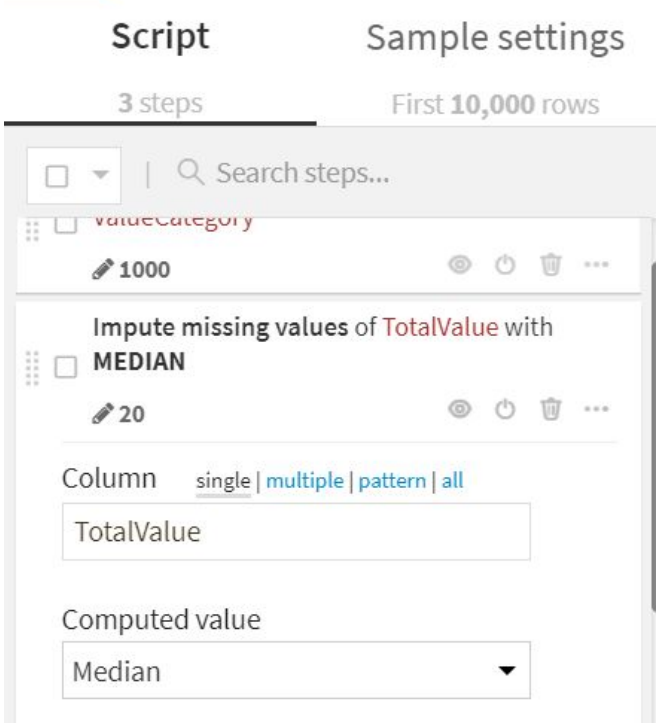
Dataiku - Missing Value Handling

Click the analyze section on the specific column name for missing values inspection



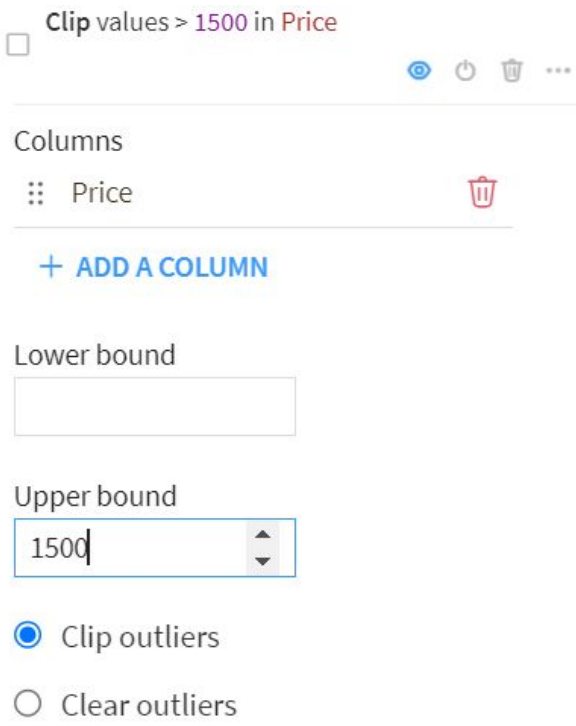


Dataiku - Missing Value Handling



- For example, impute the missing value record on the **TotalValue** column with the median value
- There are other options to impute the missing value either by mean, median, or mode

Dataiku - Anomalies and Outlier Handling



In the script section, there are two choices available,

1. Clip outliers / anomalies record by a certain value
2. Clear / remove the outliers / anomalies record if it’s exceeding the threshold

Dataiku - Data Type Inspection and Transformation

tity	Price	OrderDate	Name	Age	Gender	Location	ProductName	Category
	string	string	string	string	string	string	string	string
r	Decimal	Date (unparsed)	Text	Integer	Gender	Text	Text	Text
2	414.1818294989762	2020-05-04 18:42:53	Customer_48	22	Other	City_B	Product_81	Electronics
8	147.74873927279793	2020-06-08 10:57:29	Customer_118	43	Male	City_B	Product_83	Clothing
9	239.4031566884592	2020-02-03 00:42:43	Customer_193	30	Female	City_C	Product_38	Clothing
4	414.6722037546345	2020-08-09 16:56:32	Customer_252	65	Other	City_C	Product_95	Electronics

Price	OrderDate	Name	Age	Gender	Location	ProductName	Category
float	date	string	int	string	string	string	string
Decimal	Date (unparsed)	Text	Integer	Gender	Text	Text	Text

Data Fellowship

ML Workflow in Dataiku - Feature Engineering

What is Feature Engineering?

- Feature engineering is the process of selecting, creating, and transforming features (variables) from raw data to improve the performance of machine learning models.
- It involves identifying relevant features, encoding categorical variables, scaling numerical features, creating new features through mathematical transformations, and more

Dataiku - Column Addition

Analyze Sales_Data_Example

Script | Charts | Models | DEPLOY SCRIPT | ACTIONS

Script | Sample settings

1 step | First 10,000 rows

Search steps...

Create if, then, else statements for ValueCategory

1000

If (TotalValue >= 1000)

Then (ValueCategory= High)

Else (ValueCategory= Low)

OPEN EDITOR PANEL

+ ADD A NEW STEP

ADD A GROUP

Create if, then, else statements

CLOSE | PREVIEW | APPLY

If TotalValue >= value 1000

+ ADD A CONDITION

Then ValueCategory =(Value) High

+ ADD ELSE IF GROUP

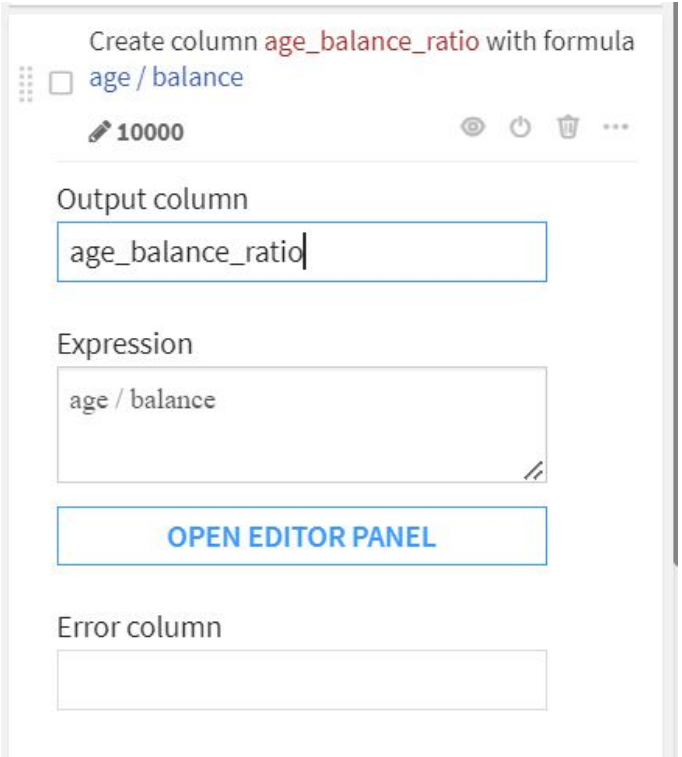
Else ValueCategory =(Value) Low

1,000 rows - 17 columns

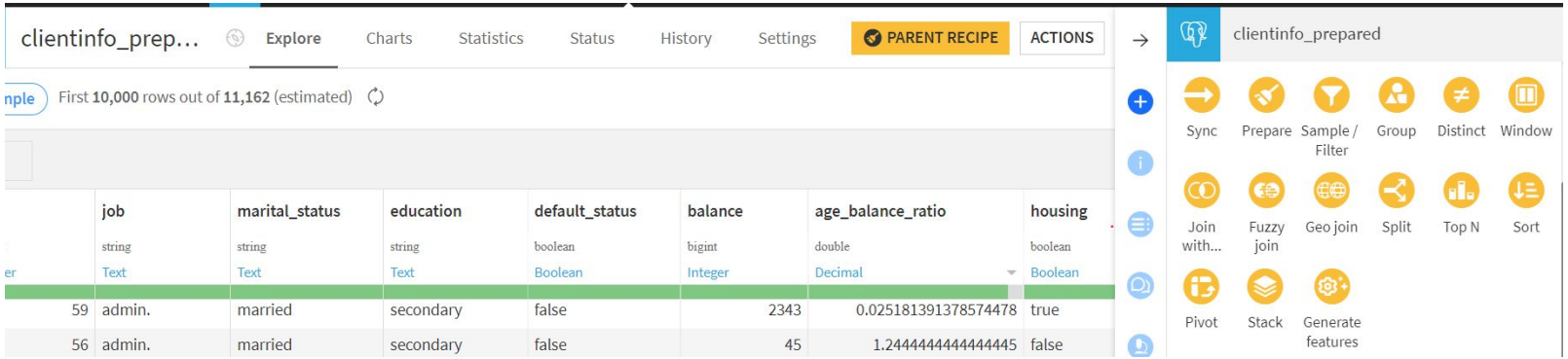
cation	ProductName	Category	UnitPrice	Month	TotalValue	ValueCategory	OrderYear
	Text	Text	Decimal	Date (unparsed)	Decimal	Text	Integer
_A	Product_37	Clothing	459.53973126783546	2020-10	1378.6191938035063	High	2020
_B	Product_81	Electronics	271.52329523860686	2020-05	543.0465904772137	Low	2020

Dataiku - Math Formula

- For the new column creation based on the math formula, it's possible to be done through dataiku with a usual math operator as shown in the example.
- For eg, to create a age_balance_ratio, age of the customer is divided by the balance they hold in their account.



Automated Feature Engineering in Dataiku



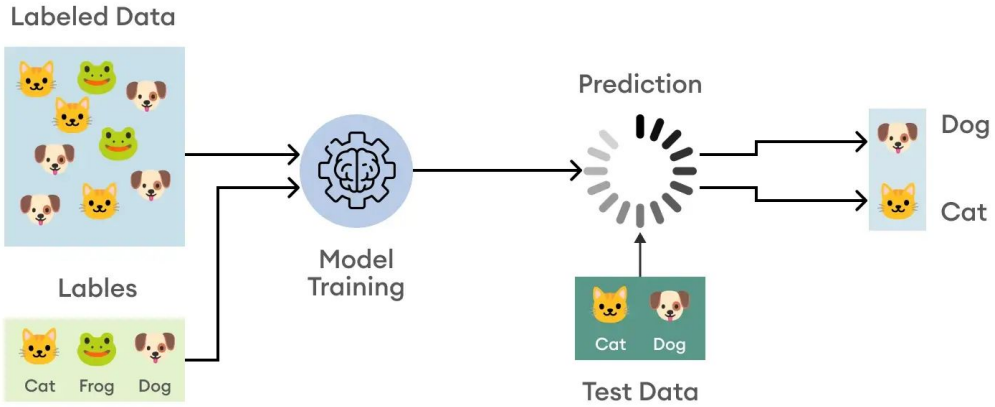
There is an automated feature engineering creation can be done through dataiku by selecting an action bar and click the **Generate features** when on the dataset menu.

Data Fellowship

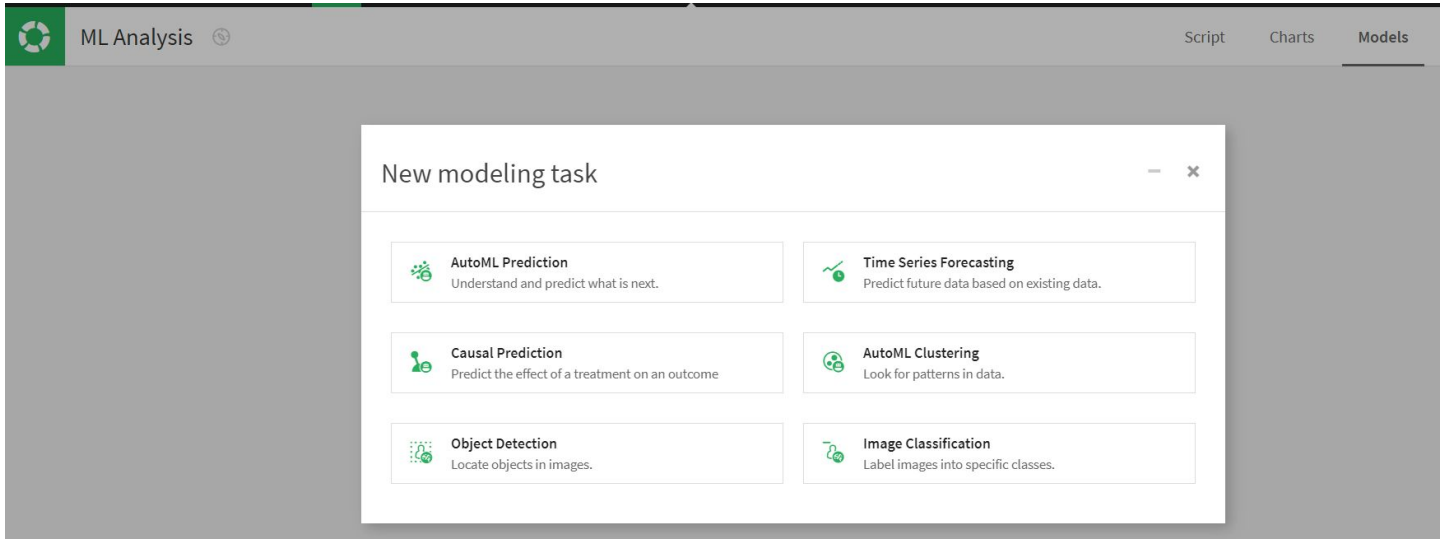
ML Workflow in Dataiku - Model Building

What is Model Building?

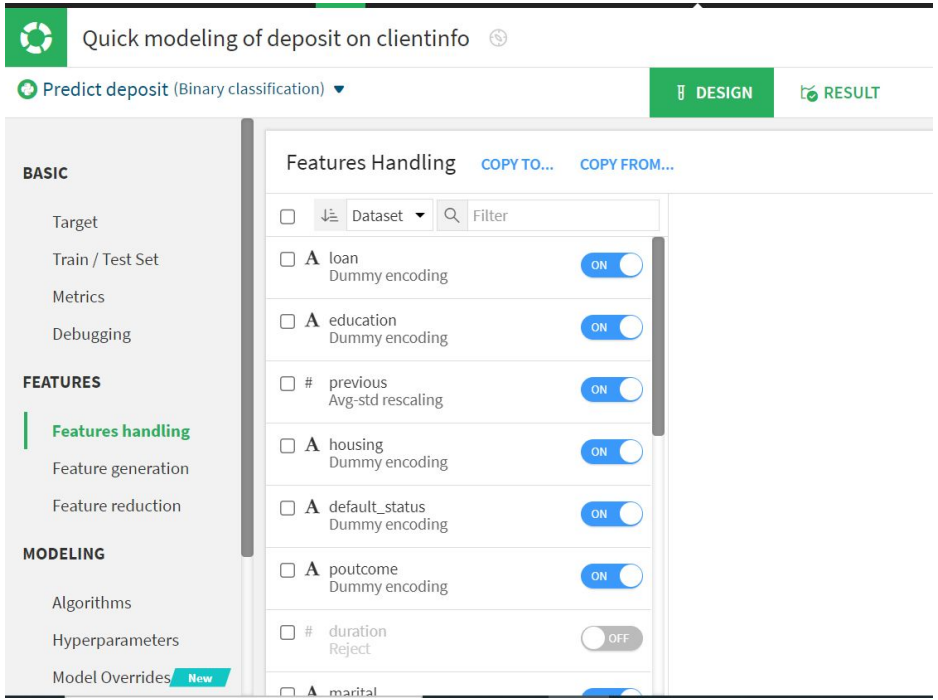
It refers to the process of creating and training a **predictive model** using a dataset to make accurate predictions or decisions based on new, unseen data. It is a crucial step in where algorithms are applied to learn patterns and relationships within the data, allowing the model to generalize and make predictions on new data instances.



Process of Machine Learning Model Creation in Dataiku



Model Building Menu Information



- **Basic**: Information / adjustment related to the overall model building such as the target variable, train-test-split, and the model metrics
- **Features**: Adjustment to handle the feature such as feature selection or feature engineering
- **Modeling:** Adjustment to handle the model / algorithm during model creation process

Data Fellowship

ML Workflow in Dataiku - Model Evaluation

What is Model Evaluation?

- Model evaluation is a critical step in the machine learning process to assess the performance and generalization capability of trained models.
- It helps determine how well a model is likely to perform on unseen data and ensures the reliability of model predictions in real-world applications.
- One of the robust model evaluation technique is utilizing **cross validation**

4-fold validation (k=4)

Data Fellowship

ML Workflow in Dataiku - Model Deployment

Deployment Options - Dataiku

1. On-Premises Deployment:

With on-premises deployment, organizations have full control over security, compliance, and scalability, ensuring data governance and regulatory compliance

2. Cloud Deployment:

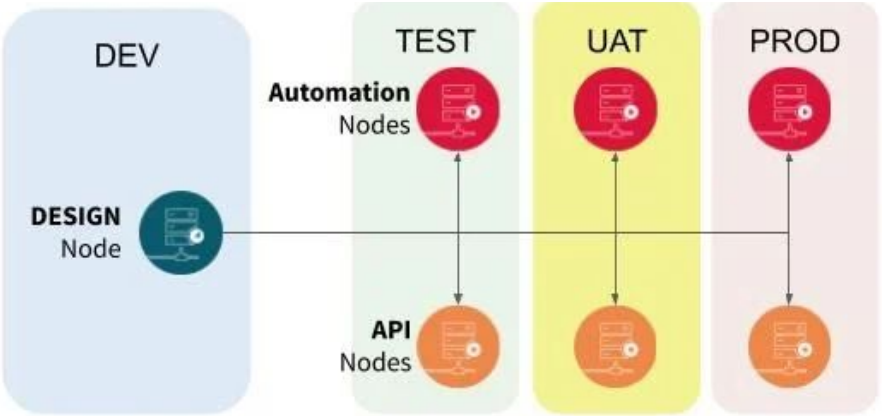
Dataiku provides seamless integration with leading cloud platforms such as AWS, Azure, and Google Cloud, allowing organizations to deploy and scale their data projects in the cloud.

Model Deployment Best Practices

1. **Continuous Integration and Deployment (CI/CD):**

Implementing CI/CD pipelines for model deployment helps streamline the deployment process, automate testing, and ensure consistency and reliability across deployments.

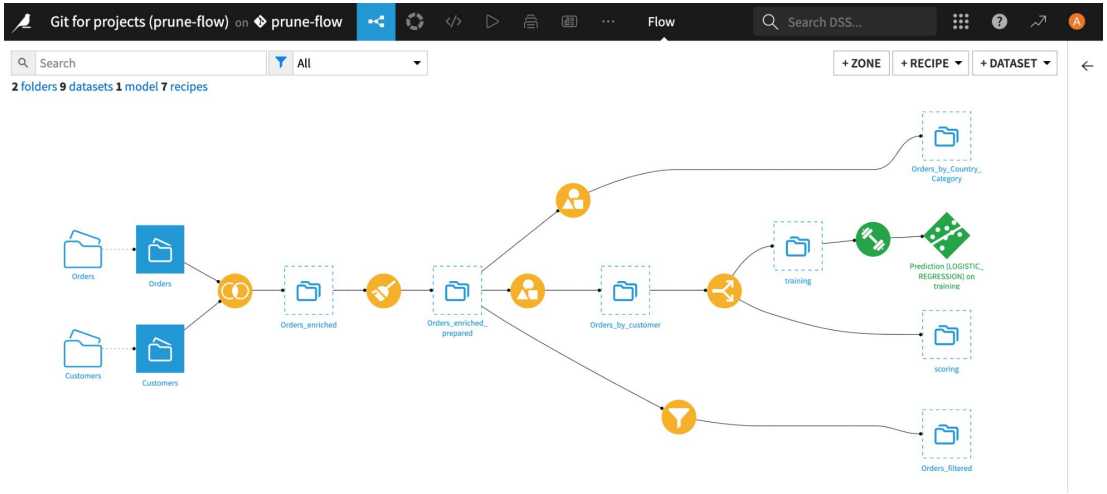
Dataiku supports integration with CI/CD tools and platforms, enabling automated deployment of models from development to production environments.



Model Deployment Best Practices

2. Version Control

Version control is essential for managing changes to models, code, and configuration files throughout the deployment lifecycle. Dataiku provides built-in version control capabilities, allowing users to track changes, collaborate effectively, and revert to previous versions if necessary



Data Fellowship

ML Workflow in Dataiku - Hands On

ML Workflow in Dataiku - Hands On

Let’s head to our dataiku and create our first model !

iykra Data Fellowship

Thank you!