Exploring the Potential of AutoML through JadBio: A Personal Hands On Journey

Sajal Agarwal
7 min read · 6 hours ago

Introduction

In todays world the importance of automating the machine learning process cannot be emphasized enough. Platforms, like JadBio have become tools for both data scientists and enthusiasts offering a user experience, in data analysis and model development without requiring coding skills. The examination of German credit data was the project's primary point, with the goal of creating a prediction model capable of assessing credit risk competently. In this article I share my experience of completing this end-to-end machine learning project using JadBio's automated platform.

Background

The German Credit Data paints a detailed portrait of individuals and their credit risk statuses, offering a fertile ground for analytical exploration. When a bank gets a loan application, it must decide whether or not to proceed with the loan approval depending on the applicant's profile. The bank's action is related with two categories of risks.

> *"If the applicant is a **good credit risk**, i.e. is likely to repay the loan, then **not approving the loan** to the person results in a **loss of business** to the bank.*
> *If the applicant is a **bad credit risk**, i.e. is not likely to repay the loan, then **approving the loan** to the person results in a **financial loss** to the bank."*

Recognizing these credit risks is pivotal for financial institutions to make enlightened decisions on loan approvals. The dataset enfolds various attributes, including credit amount, duration of the loan, and the age of the individuals, forming a rich canvas for analysis and predictions.

Objective of the Analysis

The goal of this analysis was to build a predictive model that used German credit data to accurately estimate credit risk. By harnessing the capabilities of JadBio's automated machine learning platform, the ambition was to carve a model that could pinpoint potential credit risks with precision, offering a tool that could potentially streamline decision-making processes in the financial sector.

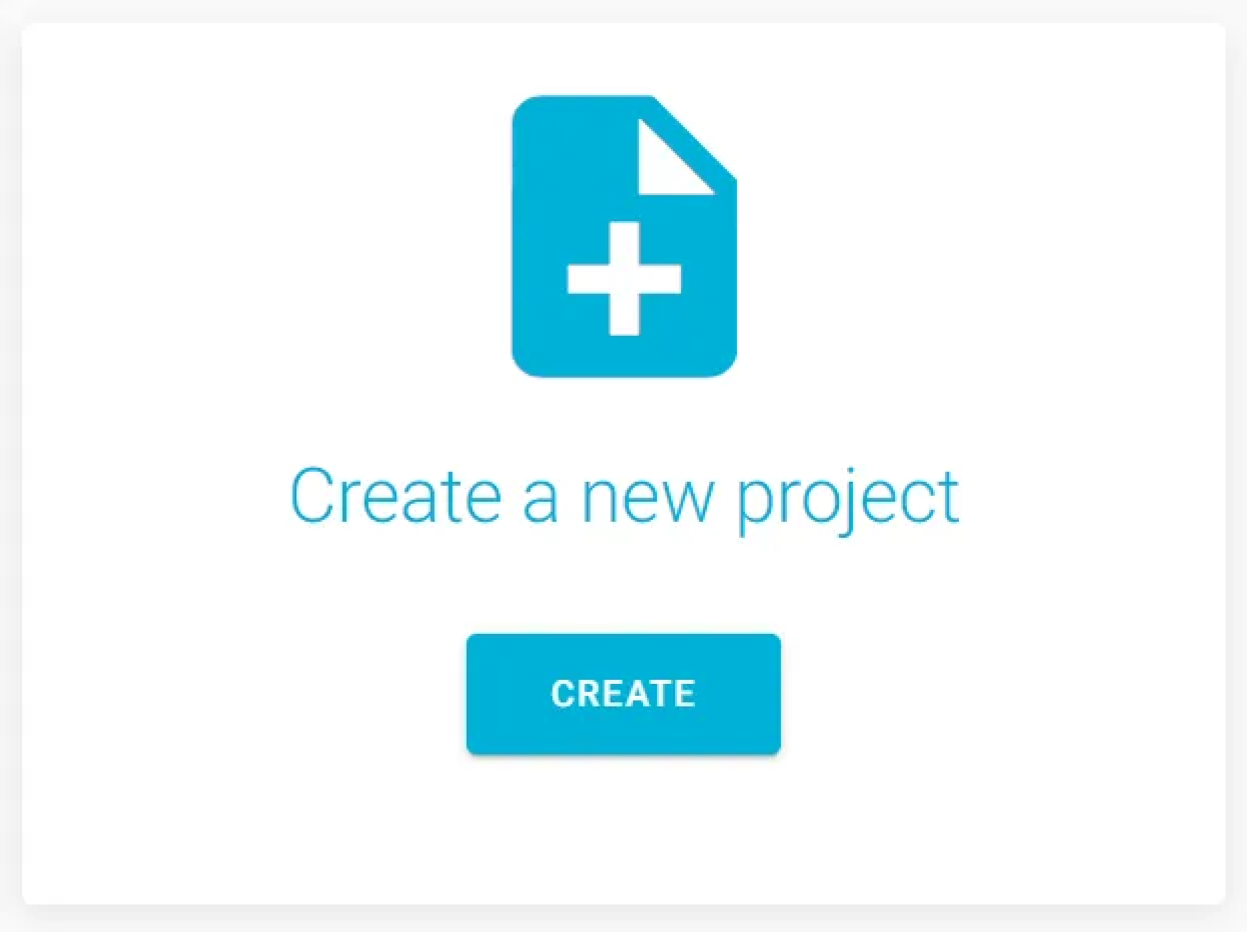
Executing the End-to-End ML Project on JadBio

1. Creating the project
2. Uploading the dataset
3. Analyze data
4. Apply model

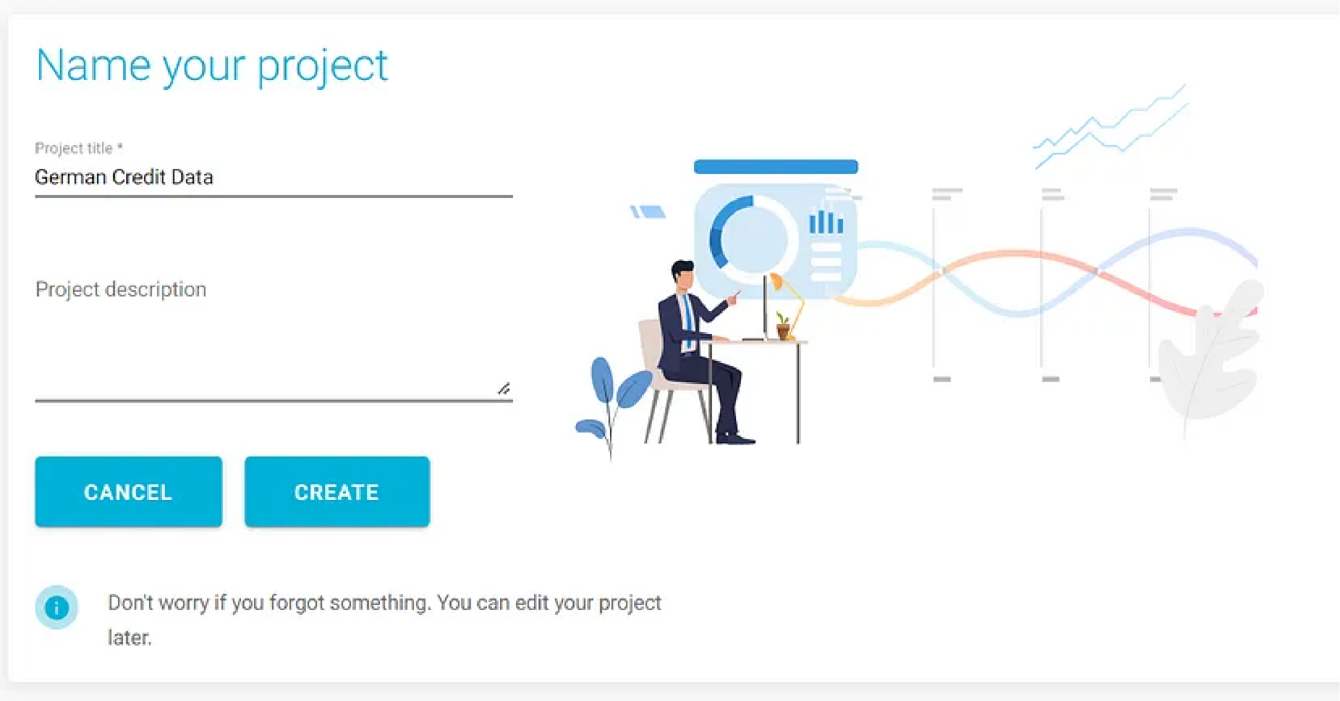
Let's dive deep into all the steps.

Creating the project

Navigate to the "Projects" section and click on "Create New Project".



Give your project a name that accurately represents your analysis.



Name your project

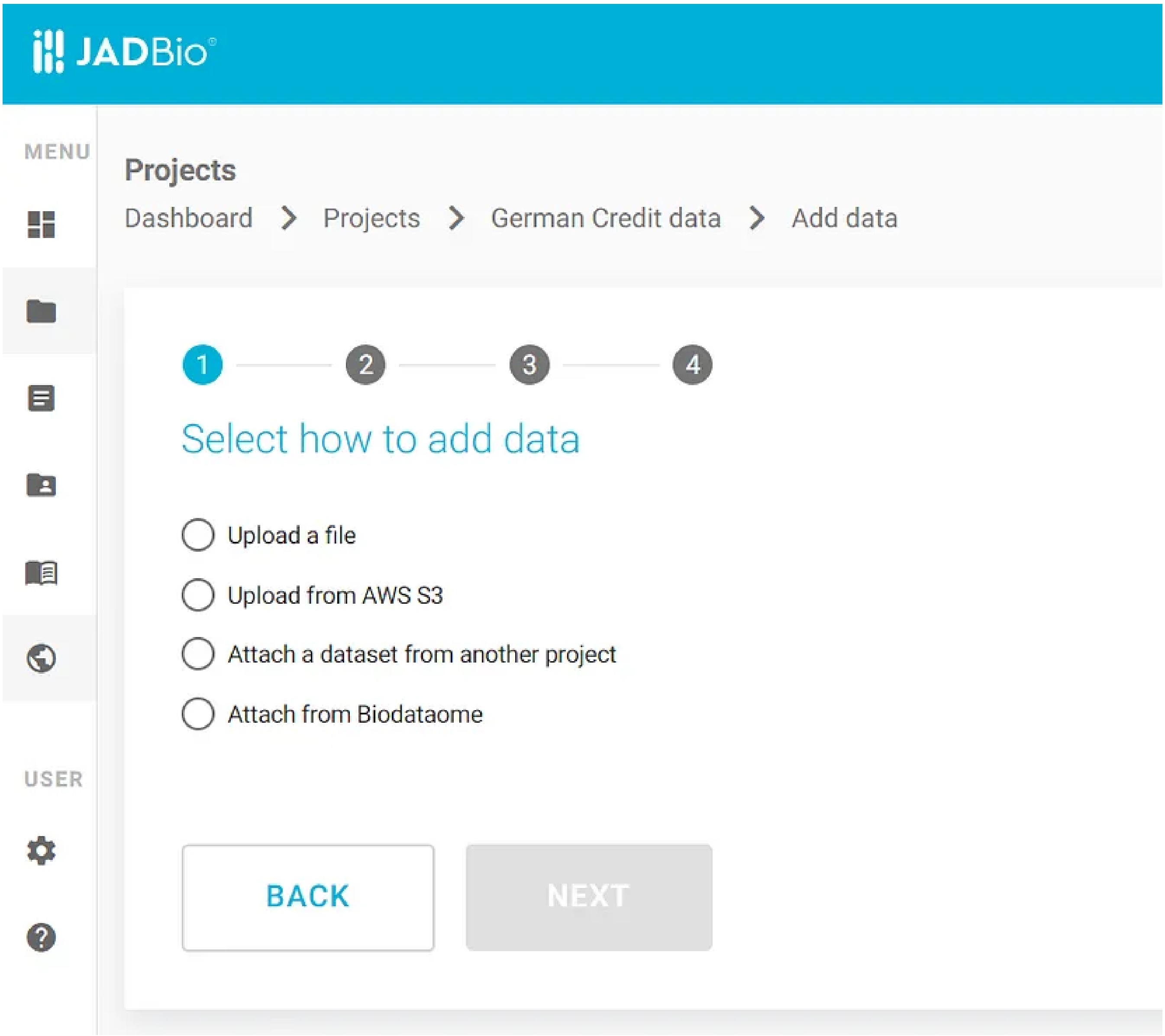
Project title *
German Credit Data

Project description

Don't worry if you forgot something. You can edit your project later.

Uploading Data

In the project, upload the dataset that you wish to analyze. For this project, the German Credit Data was chosen, a comprehensive dataset detailing credit information. You have multiple options to upload the data.



Select how to add data

○ Upload a file
○ Upload from AWS S3
○ Attach a dataset from another project
○ Attach from Biodataome

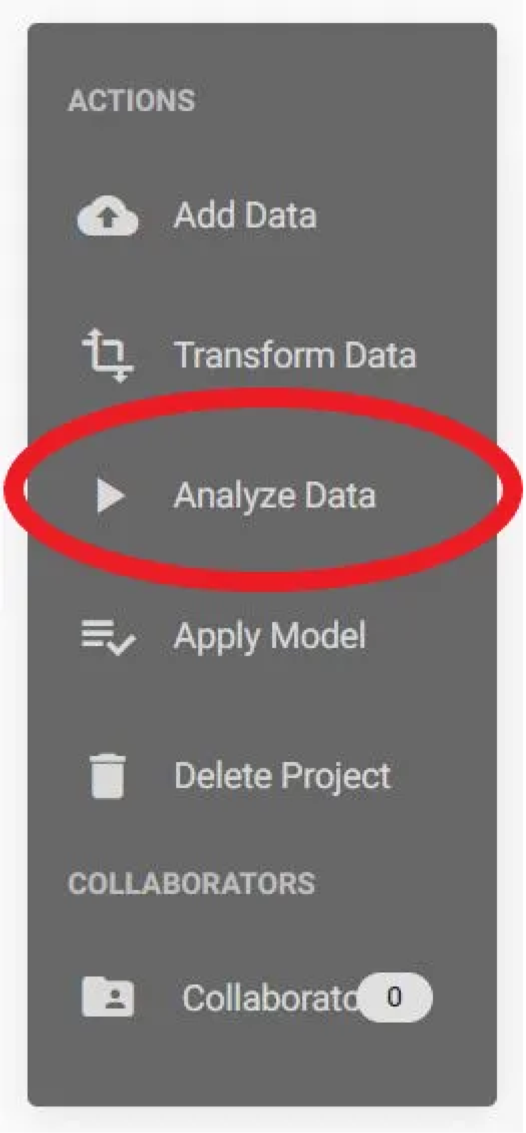
The dataset contains information about credit data with 21 columns. Here's a brief description of each column:

1. Creditability: Indicates the creditworthiness of an individual.
2. Account_Balance: Represents the current balance in the account.
3. Duration_of_Credit_monthly: Specifies the duration of the credit in months.
4. Payment_Status_of_Previous_Credit: Details the payment status of the previous credit.
5. Purpose: Indicates the purpose of the credit.
6. Credit_Amount: Specifies the amount of credit.
7. Value_Savings_Stocks: Represents the value of savings or stocks.

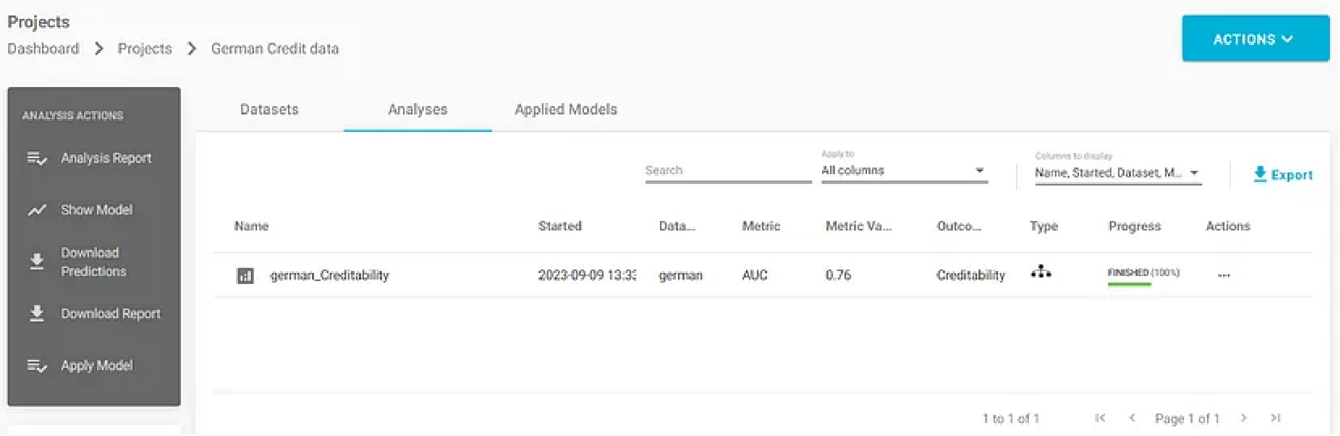
8. Length_of_current_employment: Specifies the length of current employment in years.
9. Instalment_per_cent: Indicates the percentage of installment.
10. Sex_Marital_Status: Represents the gender and marital status of the individual.
11. Guarantors: Indicates if there are any guarantors.
12. Duration_in_Current_address: Specifies the duration of stay at the current address.
13. Most_valuable_available_asset: Indicates the most valuable asset available.
14. Age_years: Specifies the age of the individual in years.
15. Concurrent_Credits: Indicates the number of concurrent credits.
16. Type_of_apartment: Represents the type of apartment.
17. No_of_Credits_at_this_Bank: Specifies the number of credits at this bank.
18. Occupation: Indicates the occupation of the individual.
19. No_of_dependents: Specifies the number of dependents.
20. Telephone: Indicates if the individual has a telephone.
21. Foreign_Worker: Specifies if the individual is a foreign worker.

Analyze the data

Once the dataset is uploaded, click on the “Analyze Data” button to initiate the analysis process.



Upon clicking the “Analyze Data” button, JadBio starts the data analysis process and generates a comprehensive report detailing the analysis results. This report is instrumental in understanding the patterns and insights derived from the data.

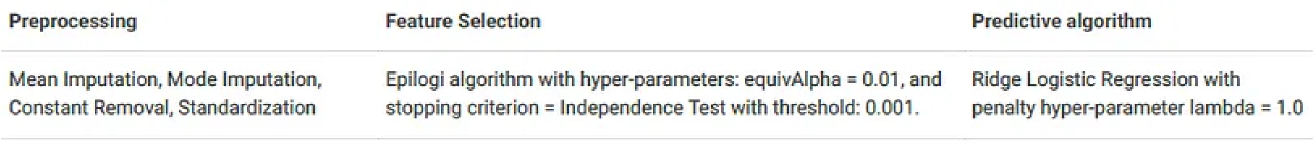


Name	Started	Data...	Metric	Metric Va...	Outco...	Type	Progress	Actions
german_Creditability	2023-09-09 13:3	german	AUC	0.76	Creditability		FINISHED (100%)	...

1 to 1 of 1 Page 1 of 1

Insights from the “Analysis Report”-

A result summary is presented for analysis optimized for Performance. The model is produced by applying the algorithms in sequence (configuration) on the training data:

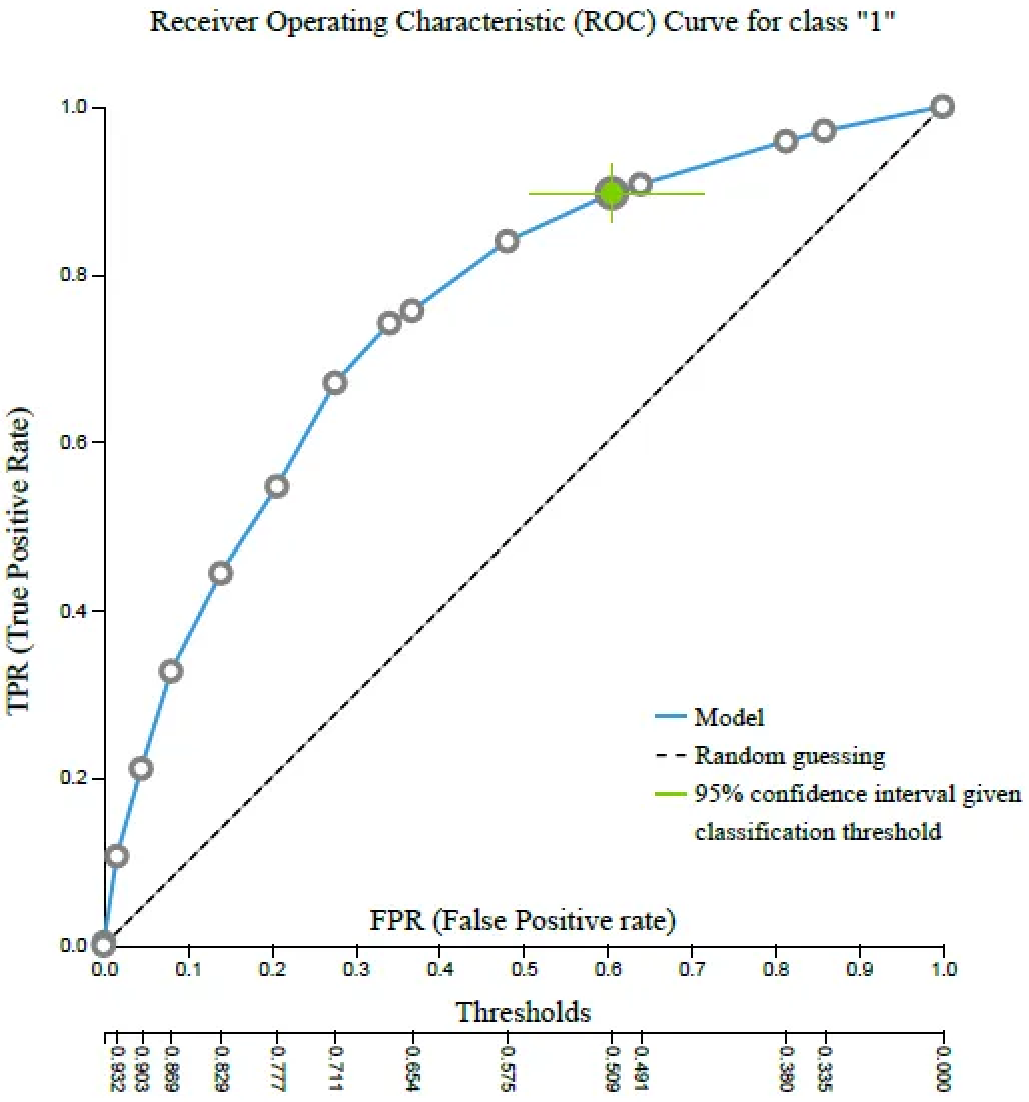


Preprocessing	Feature Selection	Predictive algorithm
Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi algorithm with hyper-parameters: equivAlpha = 0.01, and stopping criterion = Independence Test with threshold: 0.001.	Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

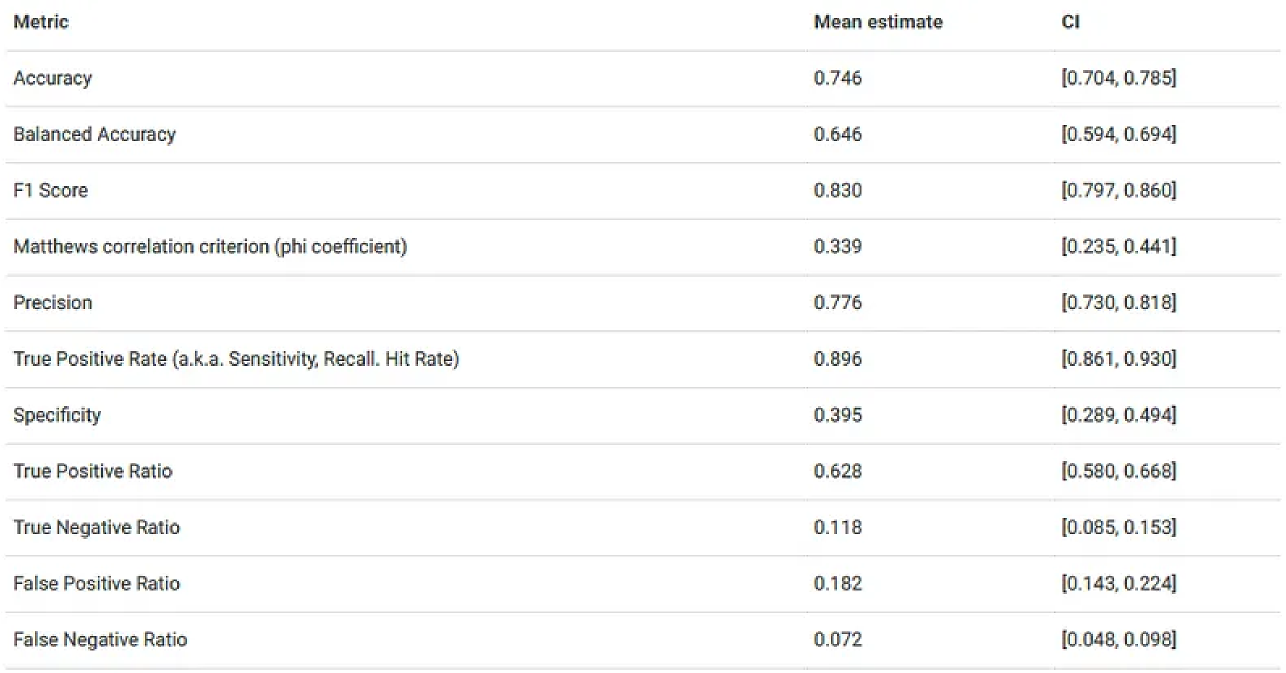
The Area Under The Curve is 0.755 with 95% confidence interval being [0.703,0.803].

The Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve) is 0.742 with 95% confidence interval being [0.695,0.788].

The Area Under the ROC Curve is shown in the figure below:



Selecting to classify as class: 1 any sample with predicted probability to be in this class above 0.5092, the model achieves:

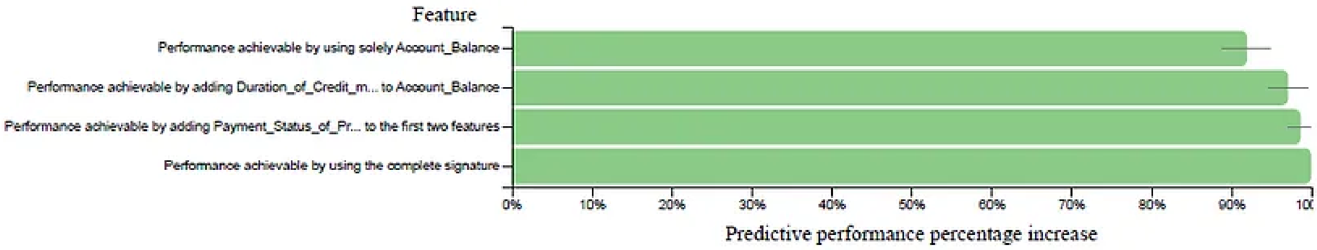


Metric	Mean estimate	CI
Accuracy	0.746	[0.704, 0.785]
Balanced Accuracy	0.646	[0.594, 0.694]
F1 Score	0.830	[0.797, 0.860]
Matthews correlation criterion (phi coefficient)	0.339	[0.235, 0.441]
Precision	0.776	[0.730, 0.818]
True Positive Rate (a.k.a. Sensitivity, Recall. Hit Rate)	0.896	[0.861, 0.930]
Specificity	0.395	[0.289, 0.494]
True Positive Ratio	0.628	[0.580, 0.668]
True Negative Ratio	0.118	[0.085, 0.153]
False Positive Ratio	0.182	[0.143, 0.224]
False Negative Ratio	0.072	[0.048, 0.098]

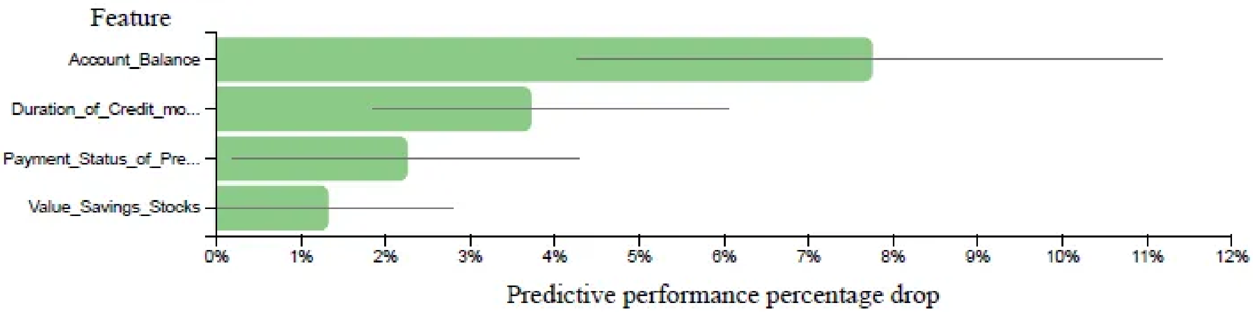
Out of the 20 features offered, four were chosen. The chosen traits include the following subset known as a signature. A single signature was found.
The system recognizes the following signatures in order of importance:
Account_Balance, Duration_of_Credit_monthly,
Payment_Status_of_Previous_Credit, Value_Savings_Stocks.

The following features cannot be substituted with others and still obtain an equal predictive performance:
Account_Balance,Duration_of_Credit_monthly,
Payment_Status_of_Previous_Credit, Value_Savings_Stocks.

The performance achieved by adding each feature in sequence to the model in comparison to the performance of the final model with all selected characteristics is shown below. The following features are included in order of importance:



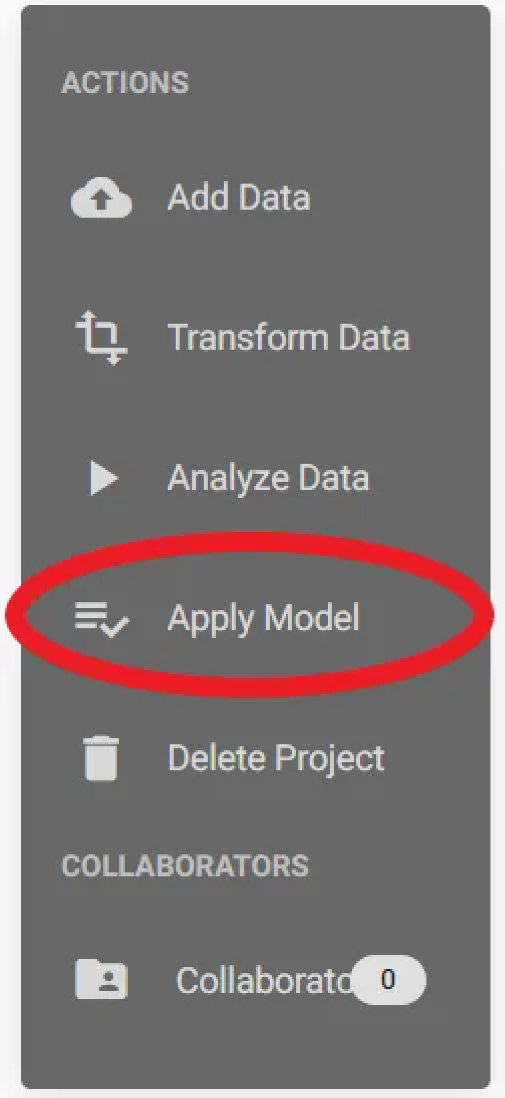
Some features may not appear to improve prediction performance in the model, yet, feature selection algorithms incorporate them in an effort to make the final model more resilient to noise. The following table compares the performance of a model with all features but one to the performance achieved when the feature is removed:



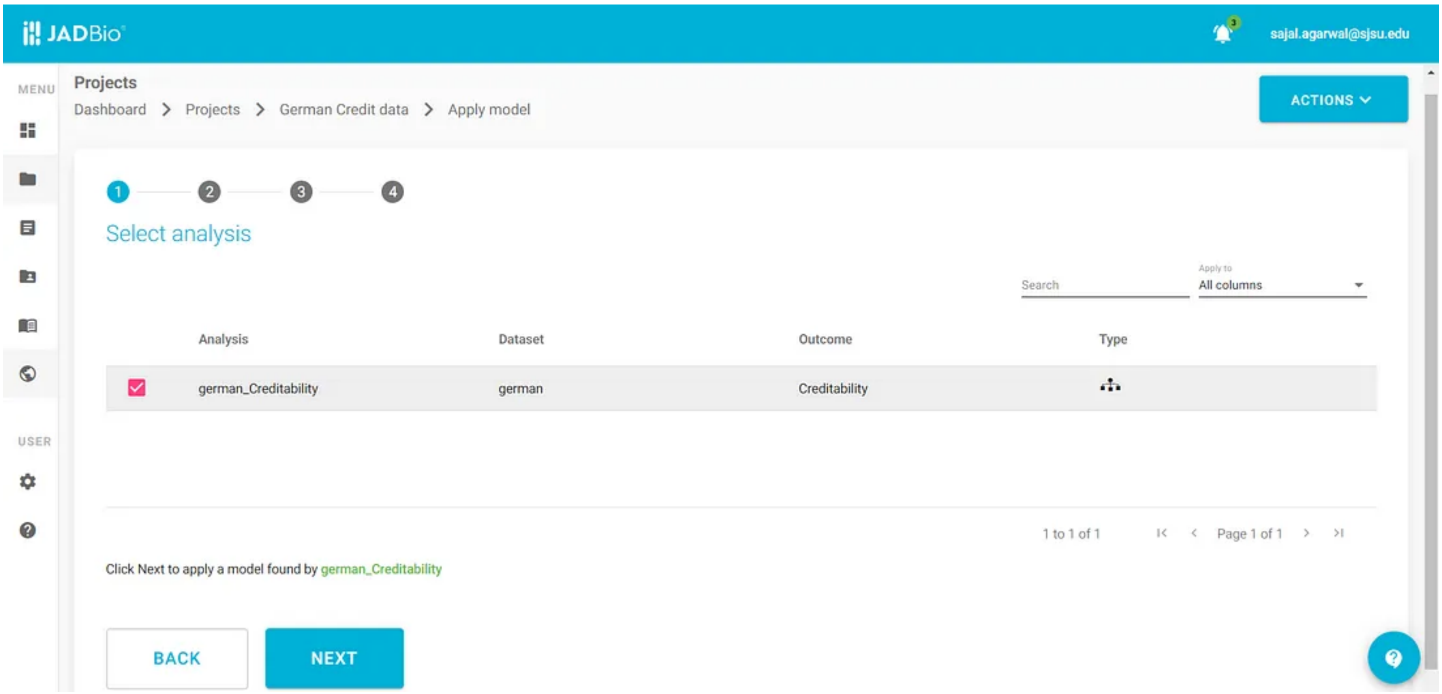
For some features there is no noticeable drop in performance when they are removed because they carry predictive information that is shared by other features selected.

Apply Model

After analyzing the data, we can go ahead with the apply model step. The user-friendly interface of the platform enabled a smooth transition from data analysis to model application, allowing for quick and efficient model validation without the necessity for coding. All we had to do is click the “Apply model” button.



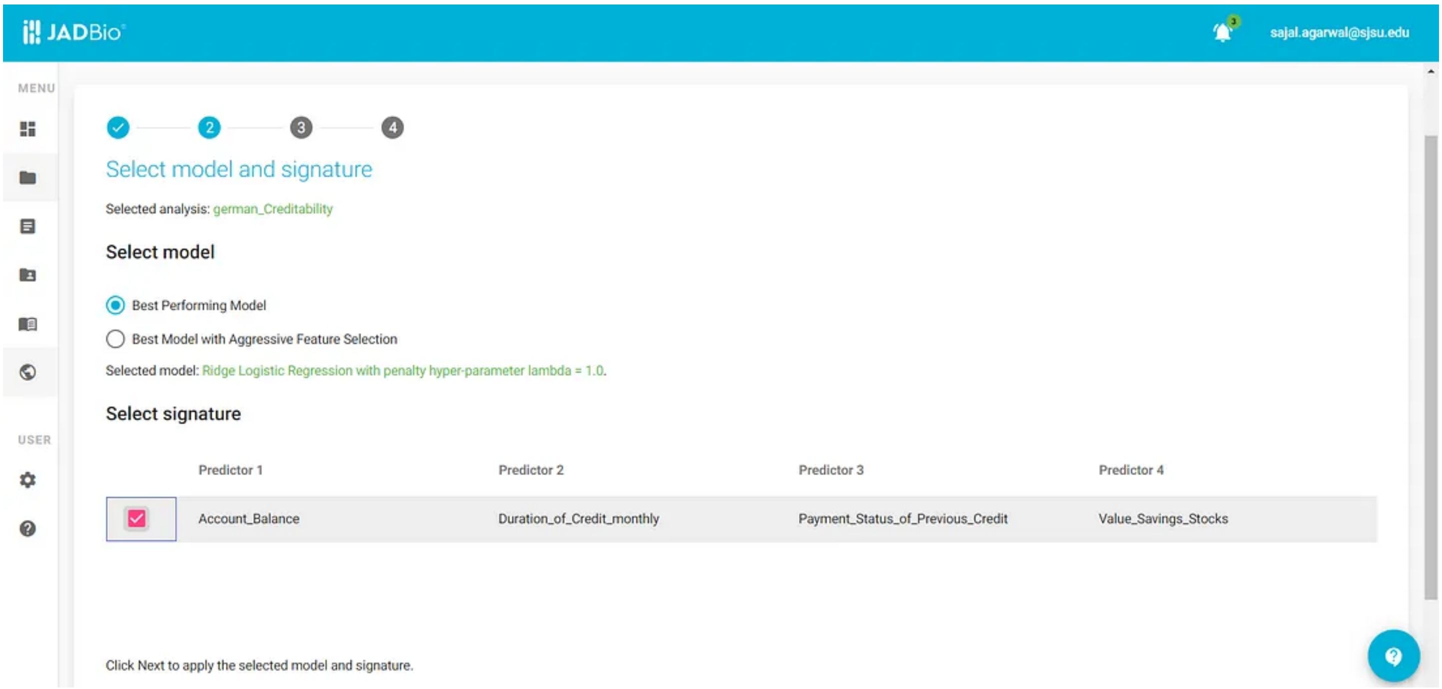
Then, we have to select the analysis to apply the model.



	Analysis	Dataset	Outcome	Type
☑	german_Creditability	german	Creditability	

Click Next to apply a model found by german_Creditability

Here we will select the model and the signature



Select model and signature

Selected analysis: german_Creditability

Select model

● Best Performing Model

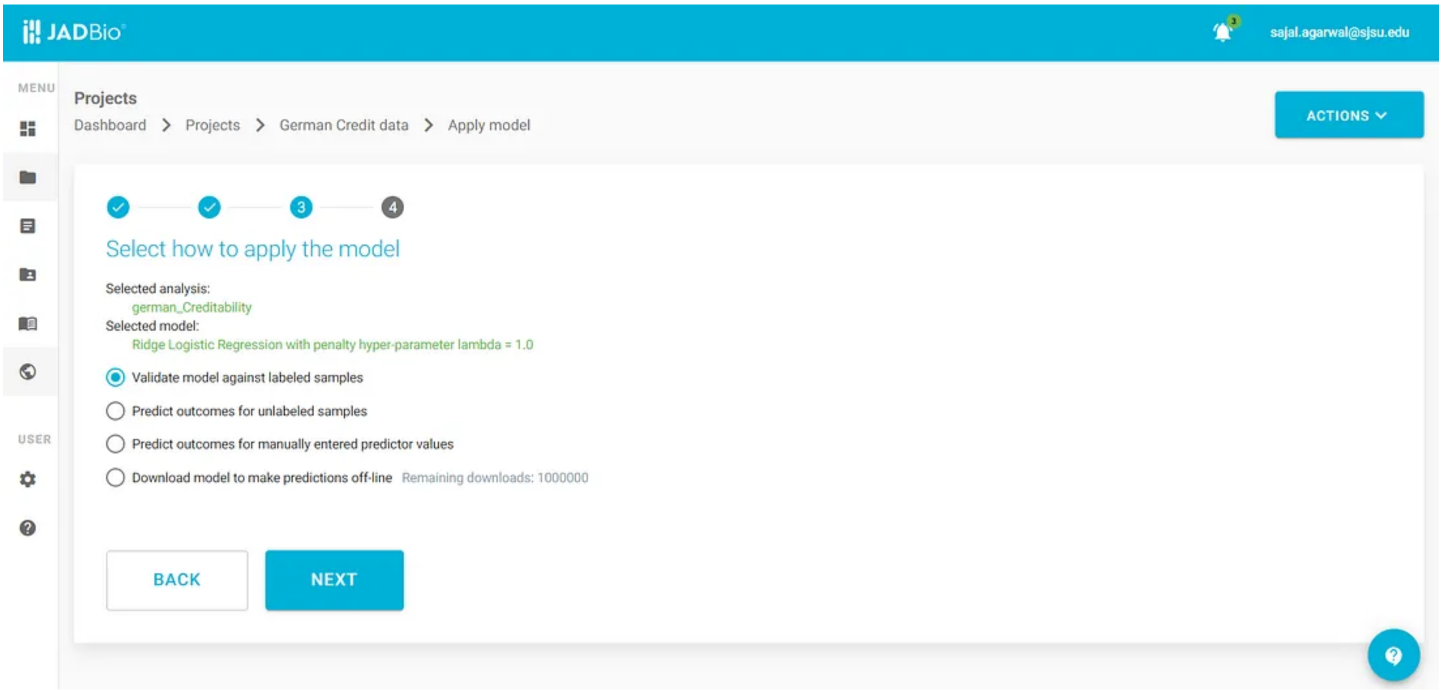
○ Best Model with Aggressive Feature Selection

Selected model: Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0.

Select signature

	Predictor 1	Predictor 2	Predictor 3	Predictor 4
☑	Account_Balance	Duration_of_Credit_monthly	Payment_Status_of_Previous_Credit	Value_Savings_Stocks

Click Next to apply the selected model and signature.



Projects

Dashboard > Projects > German Credit data > Apply model

Select how to apply the model

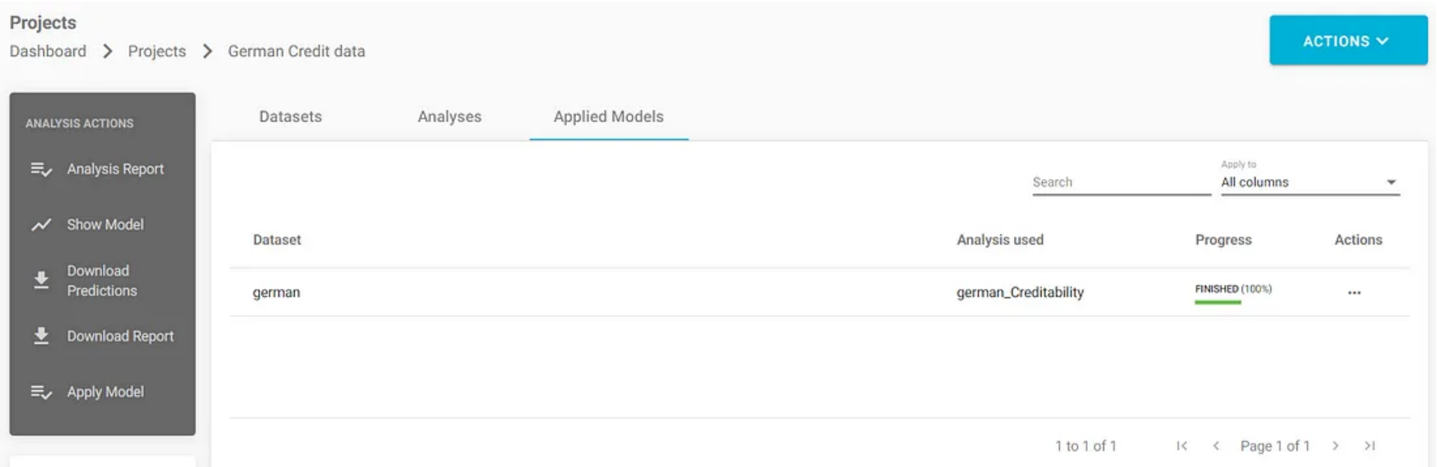
Selected analysis:
german_Creditability
Selected model:
Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

● Validate model against labeled samples

○ Predict outcomes for unlabeled samples

○ Predict outcomes for manually entered predictor values

○ Download model to make predictions off-line Remaining downloads: 1000000



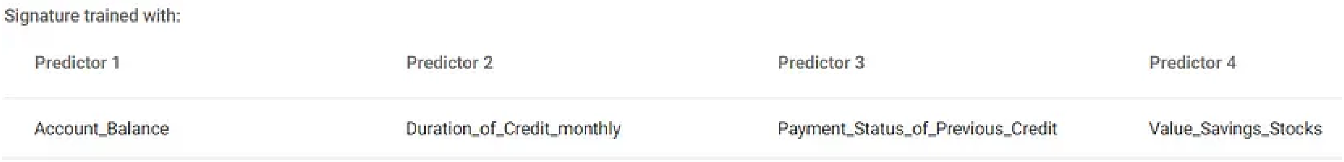
Projects

Dashboard > Projects > German Credit data

Dataset	Analysis used	Progress	Actions
german	german_Creditability	FINISHED (100%)	...

Let’s explore few insights from the results -

In the validation results page, we can see the signature it trained with.



Signature trained with:

Predictor 1	Predictor 2	Predictor 3	Predictor 4
Account_Balance	Duration_of_Credit_monthly	Payment_Status_of_Previous_Credit	Value_Savings_Stocks

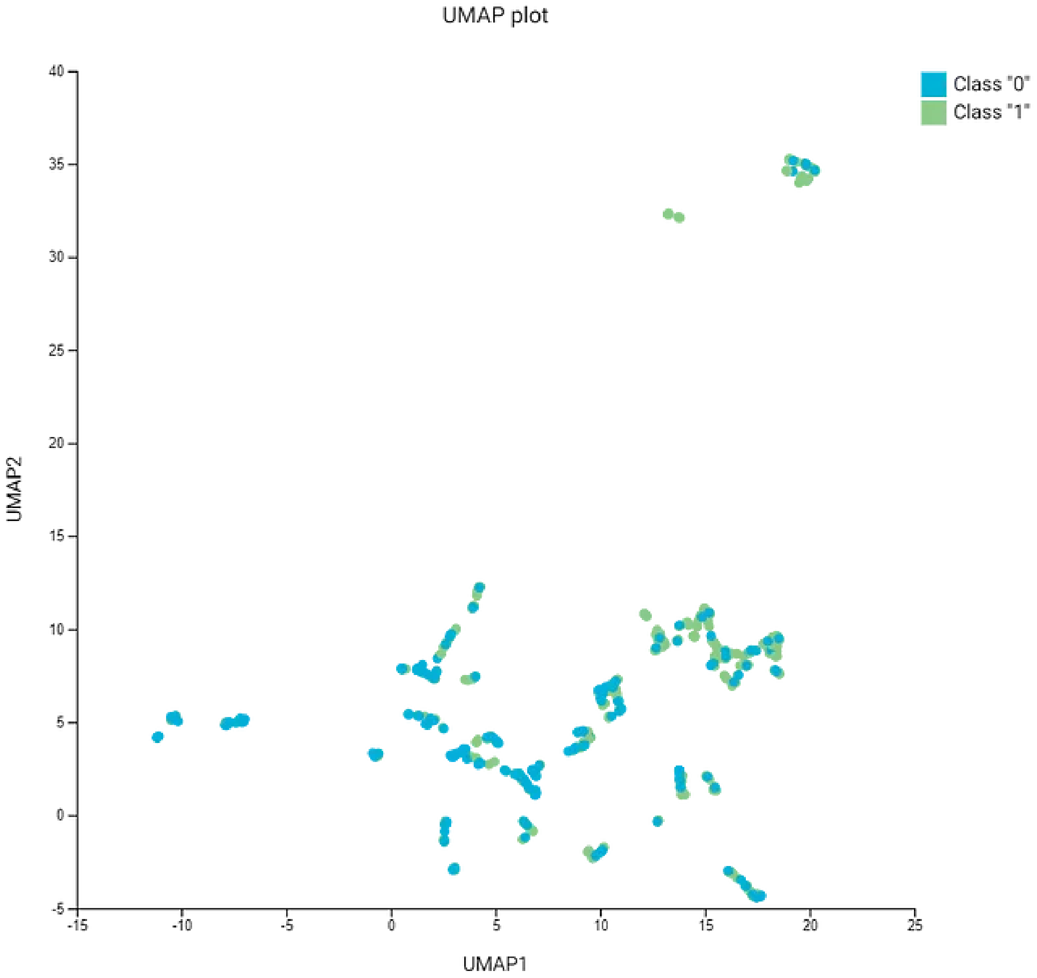
and the further results are divided into two more sections- Analysis Visualization and Performance overview.

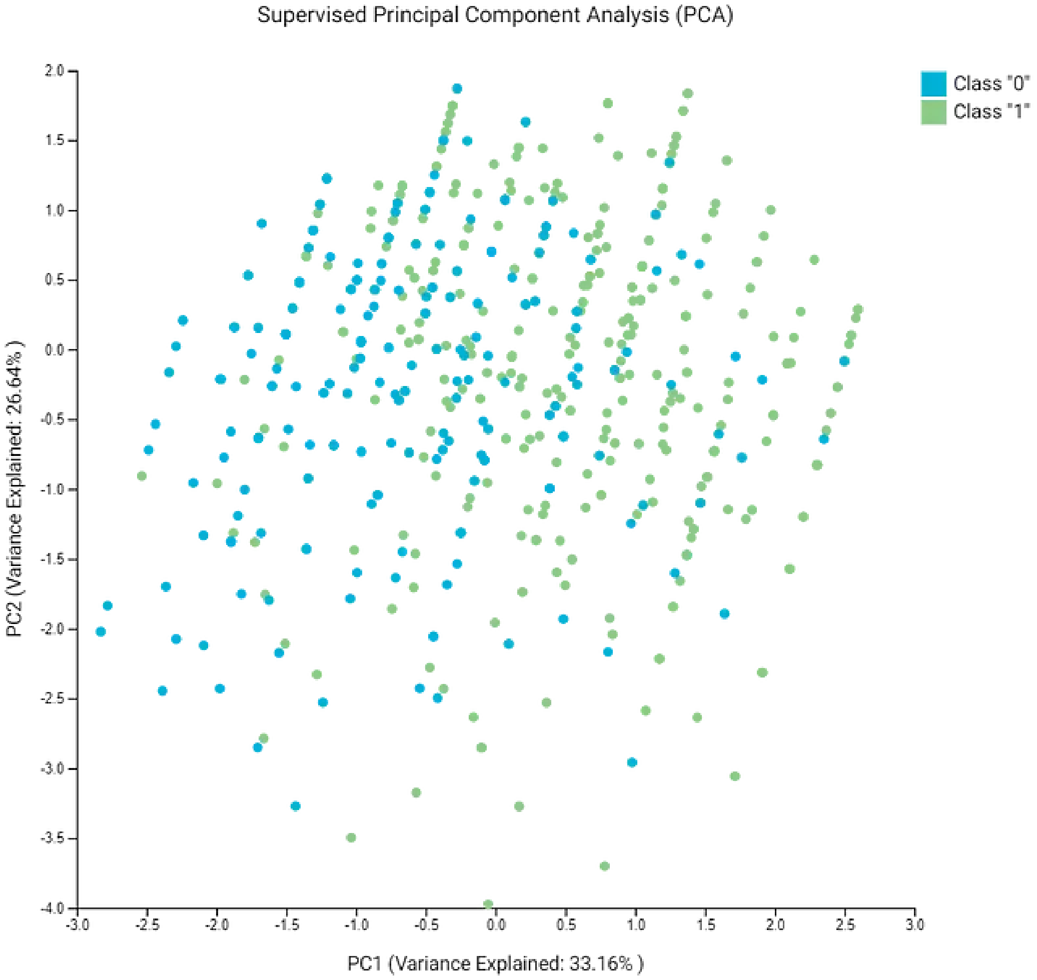
Analysis Visualization

Uniform Manifold Approximation and Projection (UMAP) attempts to learn the high-dimensional manifold on which the original data lays, and then maps it down to two dimensions. UMAP plots provides a visual aid for assessing relationships among samples.

Principal Component Analysis (PCA) is a dimensionality reduction technique that seeks the linear combinations (principal components) of the original features such that the derived features capture maximal variance.

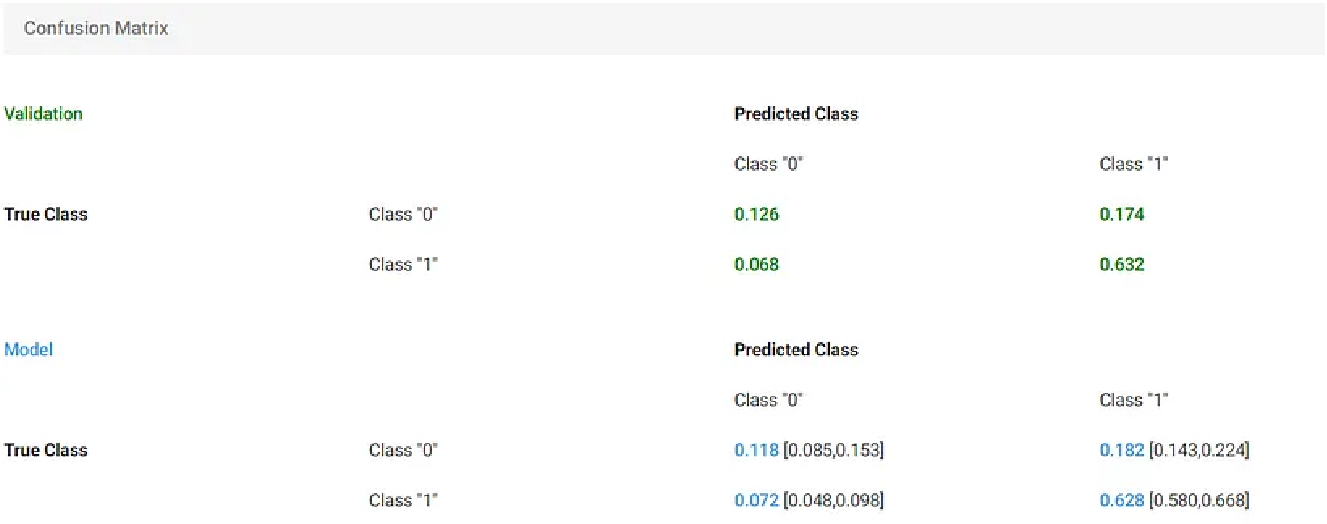
JADBio performs dimensionality reduction on a subset of the original dataset, keeping only the features included in the first signature.


UMAP plot
Class "0"
Class "1"
UMAP2
UMAP1
40
35
30
25
20
15
10
5
0
-5
-15
-10
-5
0
5
10
15
20
25


Supervised Principal Component Analysis (PCA)
Class "0"
Class "1"
PC2 (Variance Explained: 26.64%)
PC1 (Variance Explained: 33.16%)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Performance overview

A confusion matrix was generated to depict the performance of the model, illustrating the proportions of samples classified into actual and predicted classes.



Confusion Matrix

Validation		Predicted Class	
		Class "0"	Class "1"
True Class	Class "0"	0.126	0.174
	Class "1"	0.068	0.632

Model		Predicted Class	
		Class "0"	Class "1"
True Class	Class "0"	0.118 [0.085,0.153]	0.182 [0.143,0.224]
	Class "1"	0.072 [0.048,0.098]	0.628 [0.580,0.668]

Analysis of Predictions

The predictions were documented in a TXT file, capturing various details such as the probability of samples belonging to different classes and flags indicating the difficulty level of predictions. The file encapsulated the following details:

Sample Name: Denotes the identifier for each sample in the dataset.

Prob (class = 0): Indicates the probability that the sample belongs to class 0.

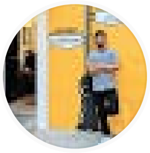
Prob (class = 1): Indicates the probability that the sample belongs to class 1.

Difficult to Predict: A flag indicating whether the prediction for the sample was challenging to ascertain.

Label: The label assigned to the sample based on the prediction.

Conclusion

In conclusion, this study demonstrates the power of automated machine learning systems in revolutionizing data science. Platforms like JadBio are paving their way in a new era of data science, characterized by efficiency, precision, and innovation, by supporting seamless processes ranging from data analysis to predictive modelling. The project’s findings offer great prospects, demonstrating a method to leveraging data science for informed and strategic decision-making in the financial sector.

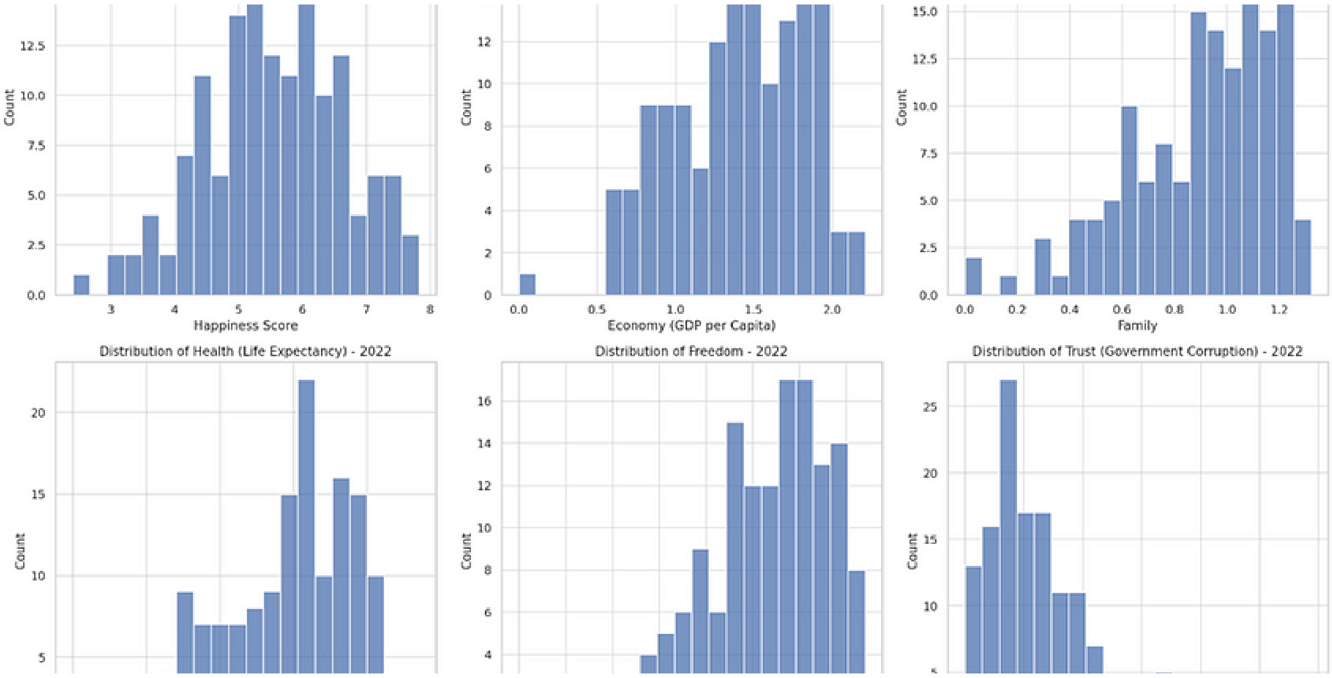


Edit profile

Written by Sajal Agarwal

1 Follower

More from Sajal Agarwal



Distribution of Health (Life Expectancy) - 2022

Distribution of Freedom - 2022

Distribution of Trust (Government Corruption) - 2022

Sajal Agarwal

Learning Data Science with ChatGPT: A Case Study on the World Happiness Report

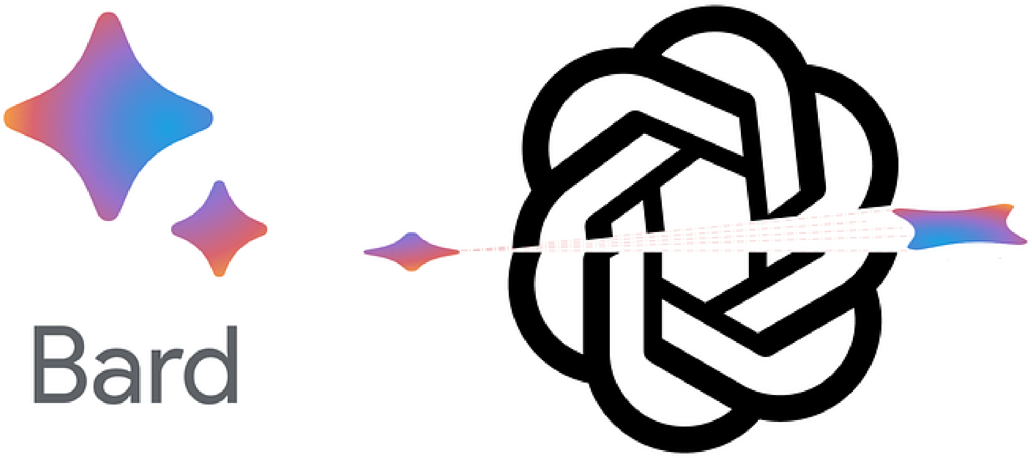
Introduction

24 min read · Aug 31

1

See all from Sajal Agarwal

Recommended from Medium



AL Anany

The ChatGPT Hype Is Over — Now Watch How Google Will Kill ChatGPT.

It never happens instantly. The business game is longer than you know.

✦ · 6 min read · Sep 1

3.8K 144

Maximilian Vogel in MLearning.ai

The ChatGPT list of lists: A collection of 3000+ prompts, examples, use-cases, tools, APIs...

Updated Sep-09, 2023. Added new prompt engineering courses, masterclasses and tutorials.

10 min read · Feb 7

7.9K 105

Lists

Staff Picks

430 stories · 265 saves

Stories to Help You Level-Up at Work

19 stories · 203 saves

Self-Improvement 101

20 stories · 537 saves

Productivity 101

20 stories · 509 saves

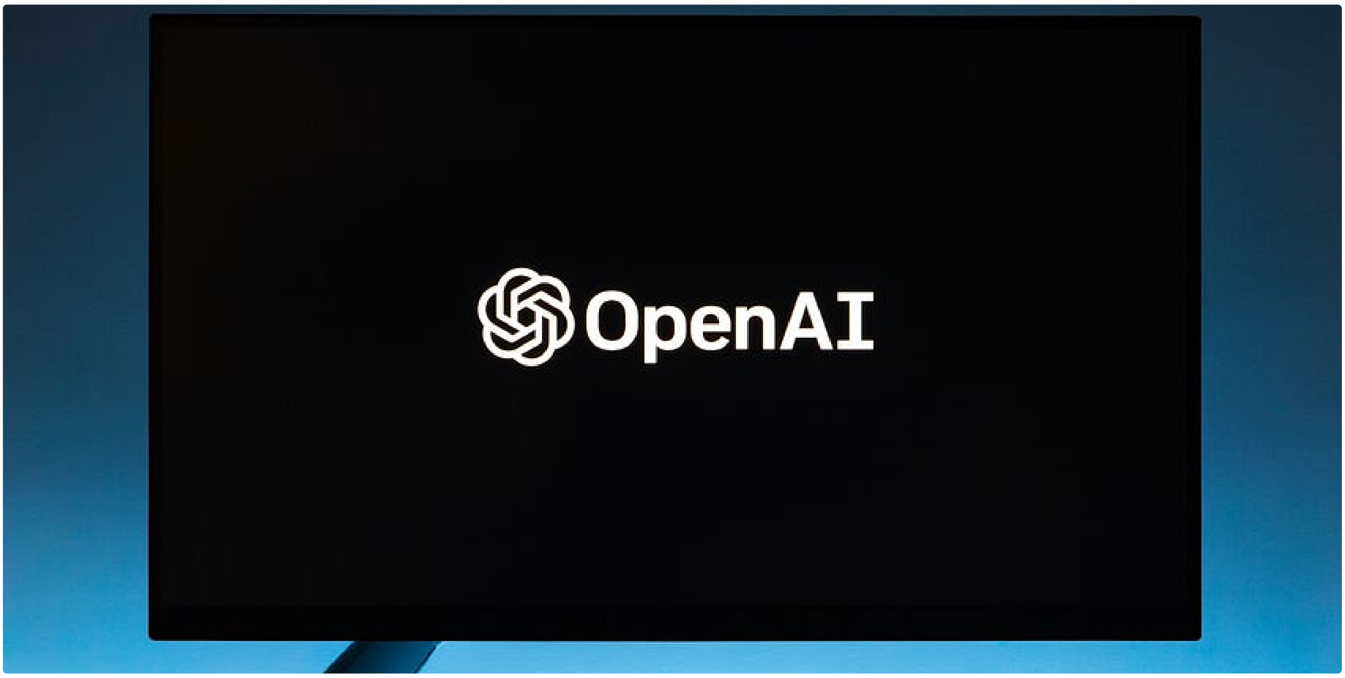
Nick Hilton

The End of the Subscription Era is Coming

You're overpaying for your porn (and journalism)

10 min read · Aug 30

5K 112



InnovatewithDataScience

Create an Generative-AI chatbot using Python and Flask: A step by step guide

Introduction

4 min read · 6 days ago

172

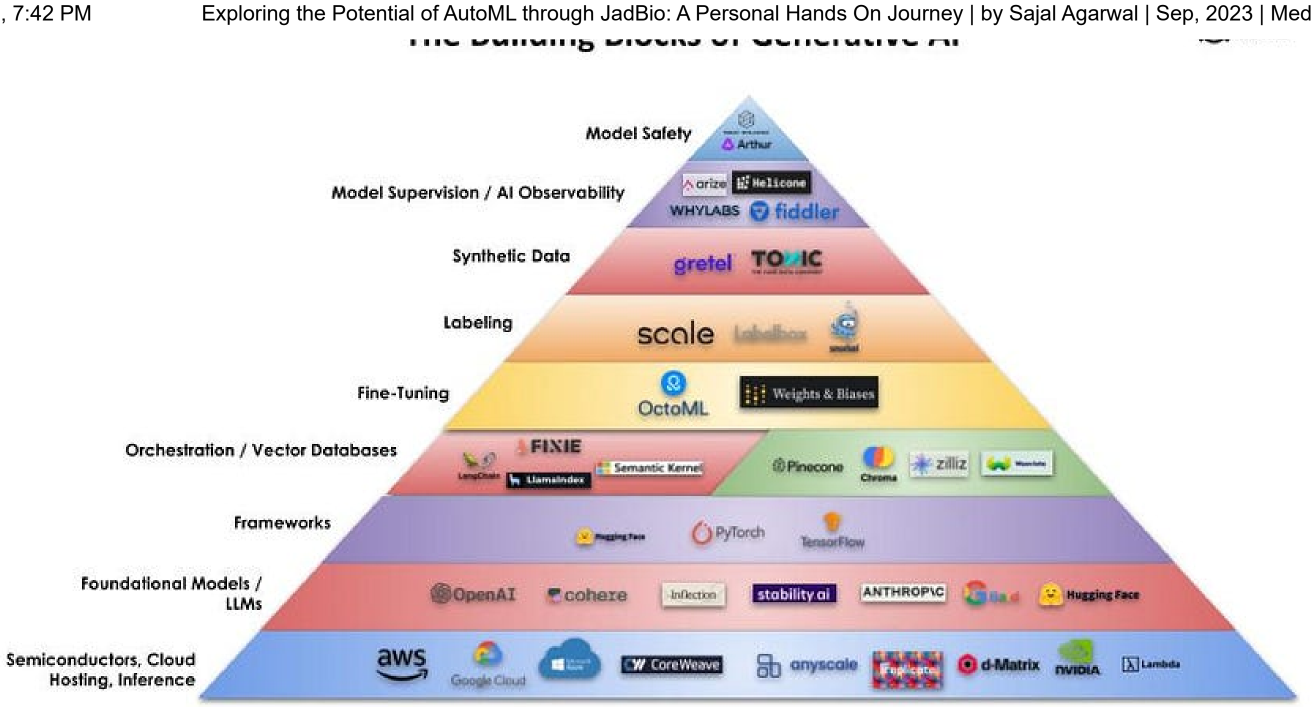
Moshe Sipper, Ph.D. in The Generator

Jailbreaking Large Language Models: If You Torture the Model Long Enough, It Will Confess!

A Cautionary Tale...

5 min read · 4 days ago

301 5



Jonathan Shriftman

The Building Blocks of Generative AI

A Beginners Guide to The Generative AI Infrastructure Stack

22 min read · Jul 10

265 3

See more recommendations