REVA UNIVERSITY
Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

Investigating Super learner for Credit Risk Modeling in Mortgage Scenario

Lalit Aggarwal

SRN: R19DM003

Date: 27th Aug 2022

PGDM/MBA in Business Analytics

Capstone Project Presentation
Year: II

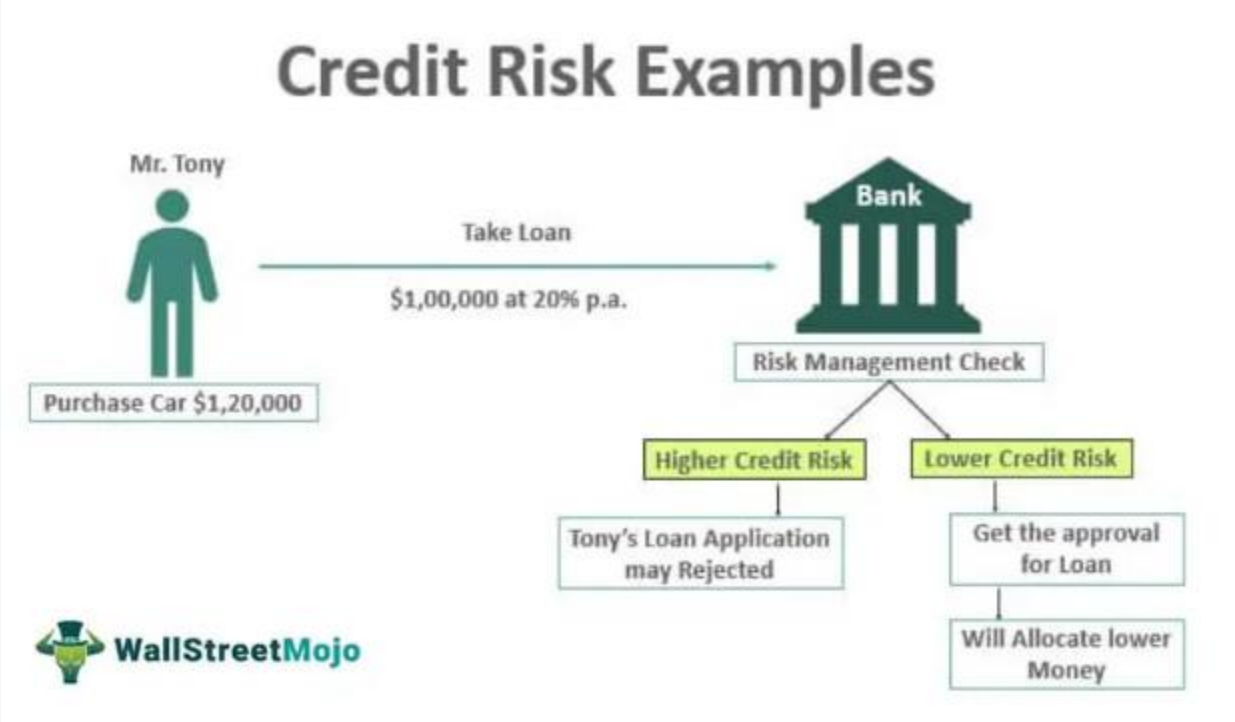
race.reva.edu.in

Agenda

Introduction

Background | Current status | Why this study

- Credit risk analysis
- Credit risk modeling
- High turnaround time Machine learning approach
- Automatic Machine learning (AutoML)



Source: www.wallstreetmojo.com

Literature Review

Sr.No.	Title	Author	Detailed Study
1	RBI CIRCULARS for all Commercial Banks	RBI Circulars, 2008-09	In this paper author describing the standard practice to be followed for handing the credit risks in the comercial banking. Stressing to **use risk expert strategies** to eliminate or minimizing the risk in lending the money.
2	Introduction to Credit Risk Modeling	Christian Bluhm, Ludger Overbeck, Christoph Wagner (2010)	In this paper author stressing **the use of credit risk modeling** where authors referred to access borrower's probability to default the loan and the impact on the lender's financial position if this default occurs.
3	Credit Risk: Modeling, Valuation and Hedging	Tomasz R. Bielecki, Marek Rutkowski (2013)	In this paper discussing about approval of loan as well **interest on the loan** based on borrower's financial status and record by the use of credit risk models. By using the latest analytics and big data tools to model credit risk. In this author **considering other factors also**, such as the development of economies and the subsequent emergence of different types of credit risk.
4	Credit Risk: Implementing Structural Models	Omomehin, Victor (2021)	In this paper author focusing the use of credit risk model to quantify **the amount of economic capital necessary to support the bank's exposures**. Author describe about structure models. Structural models are used to calculate the probability of failure of a business based on the value of its assets and liabilities. A **firm defaults if the market value of its assets is less than a debt person has to pay.**
5	Basel II: The New Basel Capital Accord, Basel Committee on Banking Supervision	Basel Committee on Banking Supervision (2003)	In this paper author discuss about **quantifying the economic capitals**. The process of allocating economic capital varies widely between banks. While some banks have implemented systems that capture most exposures across the organization, while others capture exposures within a given business line or legal entity. Besides, they have banks often developed **separate models for corporate and retail exposures**, and not all banks capture both types of exposures.

Problem Statement

Business Problem | Analytics Solution

Currently in the financial and Banking sector, they are using various Machine learning models which **takes lot of time to develop, training and hyper parameter tuning.** Usually they used to be **very complex Black Box models** where researcher used high end machine learning algorithms which are very difficult to interpret, a part of this if any customer loan couldn't approved by machine learning model, banker **couldn't provide any valid reason** why their loan application got rejected.

Project Objectives

Primary & Secondary Objectives | Expected Outcome

- Developing Different Super learner as well Base models.
- Comparison of different models.
- Explaining working of Super learners and Base models.
- Interpreting individual prediction (SHAP, PDP and ICE).

Project Methodology

Conceptual Framework | Research Design

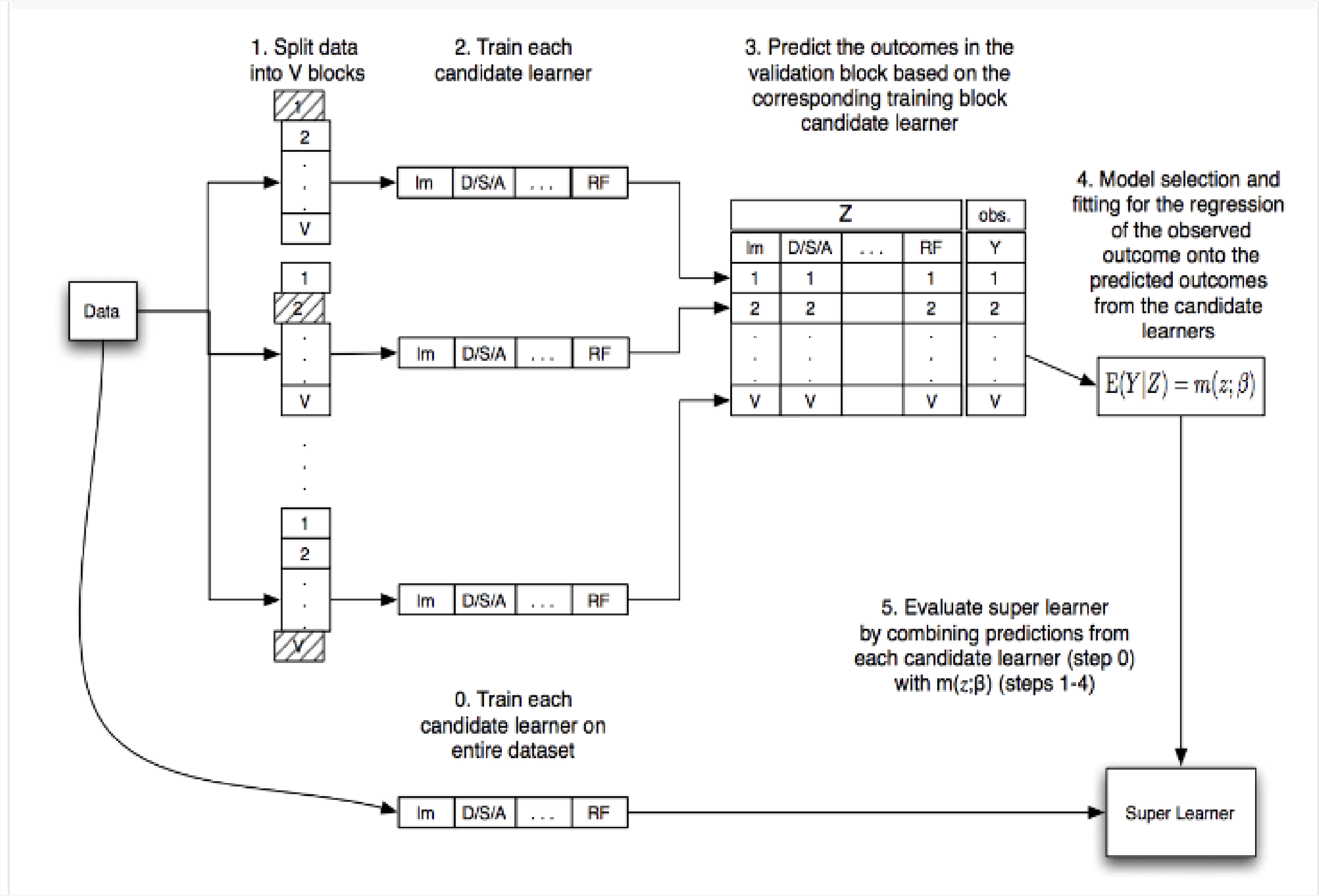


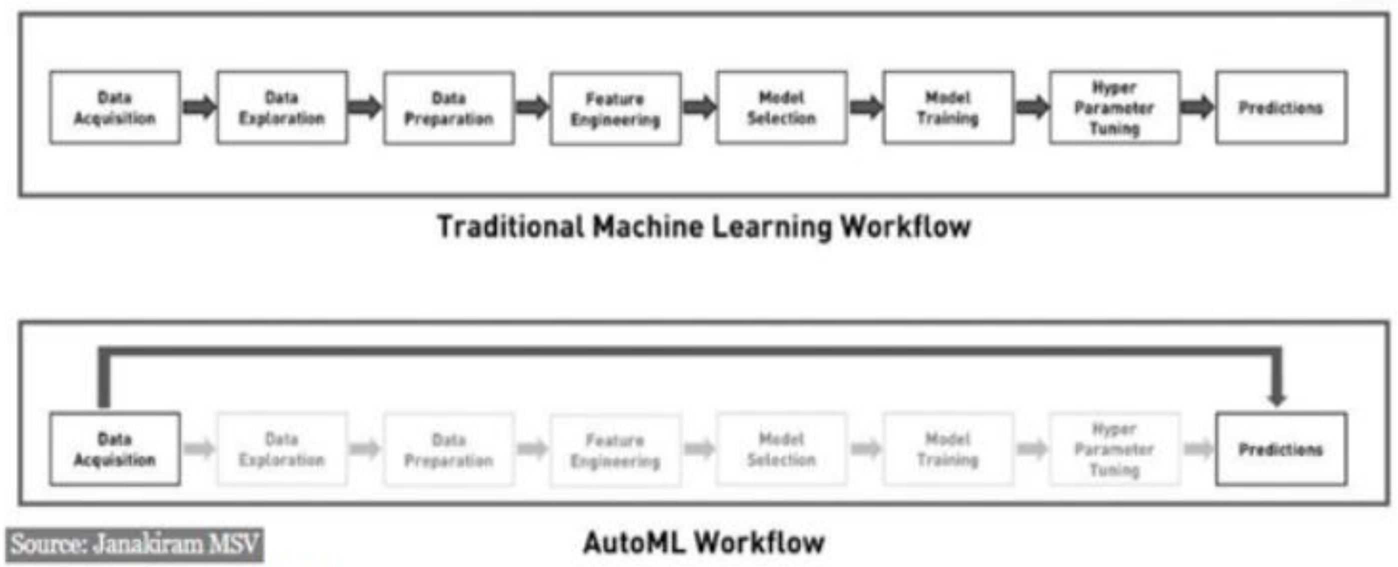
Diagram Showing the Data Flow of the Super Learner Algorithm.

The procedure can be summarized as follows:

1. Select a k-fold split of the training dataset.
2. Select m base-models or model configurations.
3. For each base model:
    a. Evaluate using k-fold cross-validation.
    b. Store all out-of-fold predictions.
    c. Fit the model on the full training dataset and store.
4. Fit a meta-model on the out-of-fold predictions.
5. Evaluate the model on a holdout dataset or use model to make predictions.

Project Methodology

Conceptual Framework | Research Design



Source: Janakiram MSV

Traditional ML vs. AutoML

Existing ML model approach vs AutoML

Process Flow

Business Understanding

Business Impact | Challenges | Monetary Impact

Impacts:

- ❖ Borrower's failure to repay a loan or meet contractual obligations.
- ❖ Interruption of cash flows and increased costs for collection.
- ❖ Properly assessing and managing credit risk can lessen the severity of a loss.

Challenges are:

- ❖ Inefficient data management.
- ❖ Getting data out of silos and into models
- ❖ Calculating Credit Risk
- ❖ Lack of Credit Risk efficient models

Biggest monetary impact has been seen in Sept. 2008, when **Lehman Brothers** meltdown, it was the 4th largest bank of USA have been in operation for 158 years. It propelled the horrors in the financial sector in USA and sparked a global financial crisis not witnessed in over last 80 years. It was involved more than **US$600 billion** in assets.

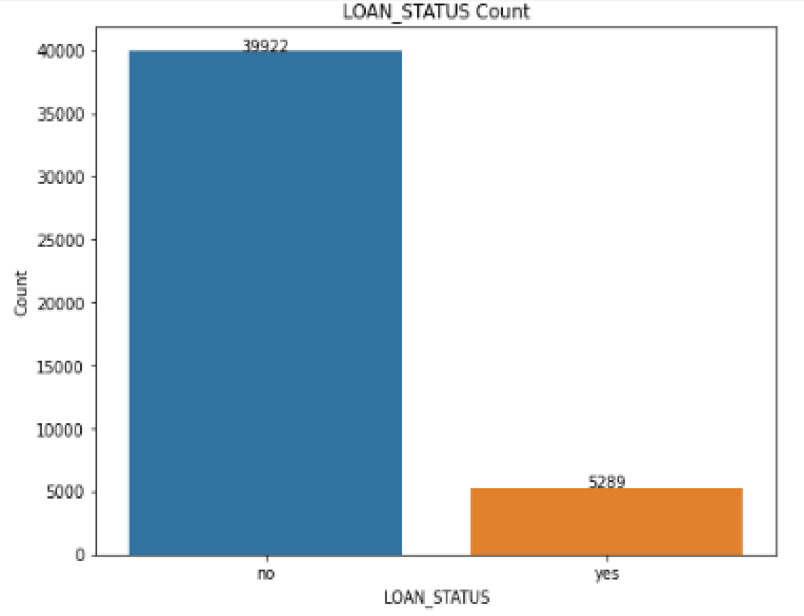
Data Understanding

Data Collection | Variables

Dataset has been collected from the internet and modified it for the Credit risk prediction as there was no relevant data available anywhere.

It had total of 9 features where Loan_Status was a target variable.

AGE	int64
JOB	object
MARITAL	object
EDUCATION	object
DEFAULT	object
HOUSING	object
LOAN	object
LOAN_STATUS	object
Income	int64



Target variable Loan_status has 45,000 of records, out of that there were 5,289 client whose loan application got approved and 39,922 client's application got rejected.

Data Preparation

Pre-processing | Techniques

	ID	AGE	JOB	MARITAL	EDUCATION
0	2836	58	management	married	tertiary
1	2837	44	technician	single	secondary
2	2838	33	entrepreneur	married	secondary
3	2839	47	blue-collar	married	unknown
4	2840	33	unknown	single	unknown

Client Info

	ID	DEFAULT	HOUSING	LOAN	Income
0	2836	no	yes	no	2194
1	2837	no	yes	no	6423
2	2838	no	yes	yes	728
3	2839	no	yes	no	2036
4	2840	no	no	no	669

Loan History

	ID	LOAN_STATUS	Unnamed: 2
0	2836	no	NaN
1	2837	no	NaN
2	2838	no	NaN
3	2839	no	NaN
4	2840	no	NaN

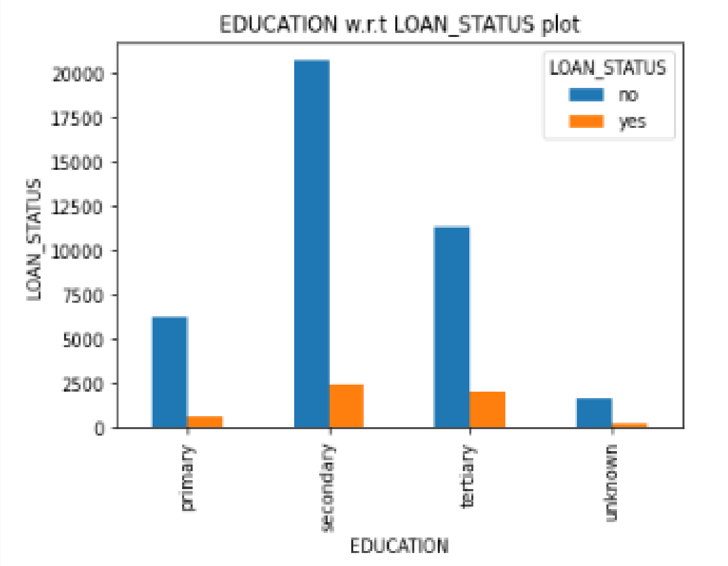
Loan Approval

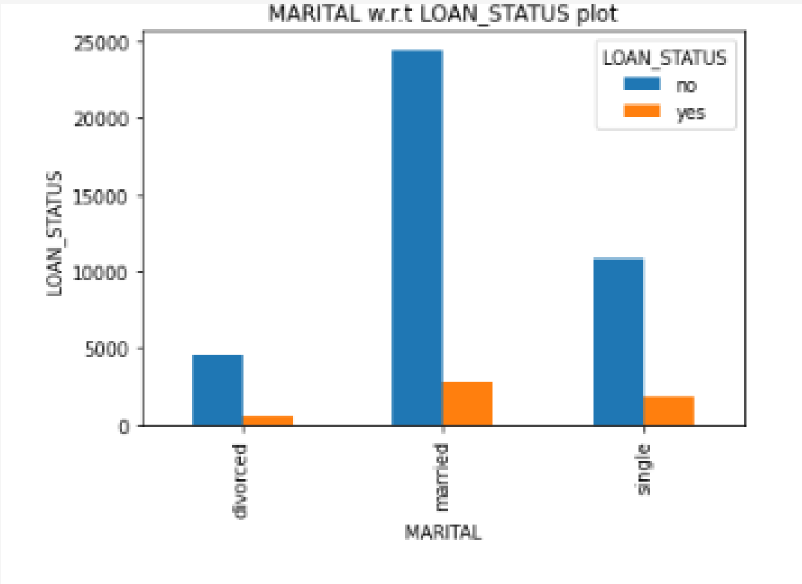
	ID	DEFAULT	HOUSING	LOAN	Income	AGE	JOB	MARITAL	EDUCATION	LOAN_STATUS
0	2836	no	yes	no	2194	58	management	married	tertiary	no
1	2837	no	yes	no	6423	44	technician	single	secondary	no
2	2838	no	yes	yes	728	33	entrepreneur	married	secondary	no
3	2839	no	yes	no	2036	47	blue-collar	married	unknown	no
4	2840	no	no	no	669	33	unknown	single	unknown	no

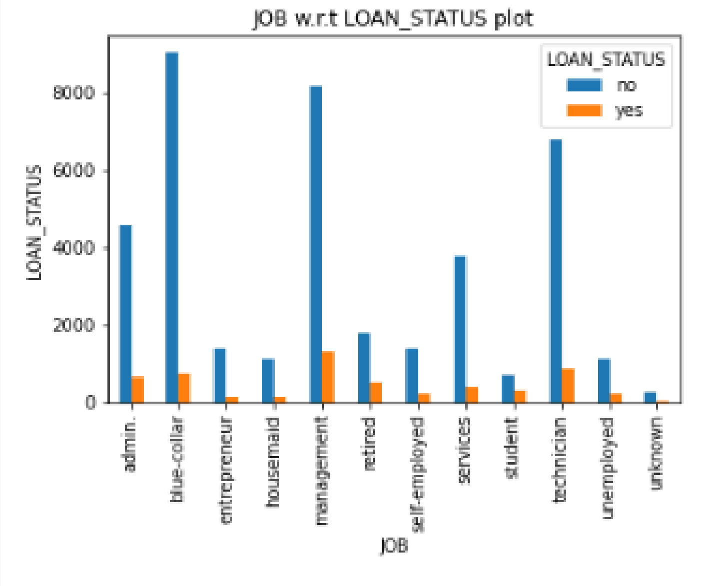
Client info + Loan History + Loan Approval = Final Data-Set

Descriptive Analytics

Multivariate Analysis | Hypothesis







Education vs Loan_Status: Clients who are **Tertiary or Secondary Education** has better chance than other of getting loan approval.

Marital vs Loan_Status: Clients who are **married or single** has better chance than divorced of getting loan approval.

Job vs Loan_Status: Clients who are working in **management role or technician** have better chance than other of getting loan approval.

Modeling

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_BestOfFamily_7_AutoML_2_20220824_183759	0.691487	0.332963	0.257057	0.363023	0.309998	0.096099
StackedEnsemble_AllModels_2_AutoML_1_20220824_181036	0.689357	0.333451	0.255477	0.355736	0.310152	0.096194
StackedEnsemble_AllModels_1_AutoML_1_20220824_181036	0.689135	0.333823	0.253464	0.360379	0.310314	0.0962945
StackedEnsemble_BestOfFamily_3_AutoML_1_20220824_181036	0.688648	0.33379	0.254494	0.36715	0.310311	0.0962927
StackedEnsemble_BestOfFamily_8_AutoML_2_20220824_183759	0.688503	0.335648	0.256424	0.35715	0.31115	0.0968144
StackedEnsemble_BestOfFamily_2_AutoML_1_20220824_181036	0.688061	0.334248	0.252488	0.36203	0.31047	0.0963914
GBM_grid_1_AutoML_2_20220824_183759_model_3	0.685723	0.335591	0.249722	0.369964	0.311378	0.0969561
StackedEnsemble_BestOfFamily_1_AutoML_1_20220824_181036	0.685462	0.335038	0.248778	0.371197	0.310938	0.0966824
XRT_2_AutoML_2_20220824_183759	0.685185	0.340844	0.253002	0.371869	0.314104	0.0986611
GBM_grid_1_AutoML_2_20220824_183759_model_6	0.684839	0.335535	0.24968	0.361034	0.311281	0.0968956
GBM_7_AutoML_2_20220824_183759	0.683234	0.336114	0.247742	0.362919	0.311328	0.0969248
GBM_2_AutoML_1_20220824_181036	0.683234	0.336114	0.247742	0.362919	0.311328	0.0969248
GBM_10_AutoML_2_20220824_183759	0.682514	0.33658	0.246377	0.361747	0.3116	0.0970945
GBM_6_AutoML_2_20220824_183759	0.682412	0.336528	0.248819	0.374916	0.31137	0.0969513
GBM_1_AutoML_1_20220824_181036	0.682412	0.336528	0.248819	0.374916	0.31137	0.0969513
GBM_5_AutoML_1_20220824_181036	0.682328	0.336295	0.246526	0.365245	0.311548	0.0970621
GBM_grid_1_AutoML_2_20220824_183759_model_7	0.682295	0.336875	0.249792	0.365376	0.311946	0.09731
GBM_3_AutoML_1_20220824_181036	0.682268	0.336884	0.247817	0.369229	0.311449	0.0970007
GBM_8_AutoML_2_20220824_183759	0.682268	0.336884	0.247817	0.369229	0.311449	0.0970007
GBM_grid_1_AutoML_2_20220824_183759_model_2	0.682258	0.335538	0.250499	0.373032	0.31113	0.0968019
DRF_2_AutoML_2_20220824_183759	0.681966	0.343633	0.249756	0.367832	0.312976	0.0979539
XRT_1_AutoML_1_20220824_181036	0.678356	0.339351	0.245942	0.36422	0.313121	0.0980448
DRF_1_AutoML_1_20220824_181036	0.676756	0.345244	0.246482	0.376885	0.313318	0.098168
GBM_4_AutoML_1_20220824_181036	0.676389	0.339594	0.242271	0.364259	0.312362	0.0975701
GBM_9_AutoML_2_20220824_183759	0.676389	0.339594	0.242271	0.364259	0.312362	0.0975701
GBM_grid_1_AutoML_2_20220824_183759_model_4	0.667997	0.341998	0.231962	0.384912	0.313895	0.0985304
GLM_2_AutoML_2_20220824_183759	0.665746	0.341551	0.216353	0.372929	0.314072	0.0986411

Leader-board of AutoML

Modeling Techniques | Modeling Process | Model Building

With the following parameters, AutoML produced 38 machine learning statistical models (Superlearner and base models)

- max_runtime_sec = 600,
- max_models = 50,
- Balance_classes = True,
- Stopping metric = AUC
- Stopping rounds = 3

It's a combination of following models:

1. StackedEnsemble_BestOfFamily
2. StackedEnsemble_AllModels
3. Base models (GBM, XRT, DRF, GLM , DeepLearning, GBM_grid, DeepLearning_grid)

- DRF : Distributed RF, XRF : Xtremely Randomize Trees, GLM : Generalized Linear Models
- Grid-searching is **the process of scanning the data to configure optimal parameters for a given model**..

Model Evaluation

Results | Interpretation | Insights

```
metalearner.varimp()
```

```
[('GBM_grid_1_AutoML_2_20220824_183759_model_3',
  0.2754656672477722,
  1.0,
  0.38459352242163053),
 ('DRF_2_AutoML_2_20220824_183759',
  0.26613906025886536,
  0.9661423977728669,
  0.37157210792034695),
 ('XRT_2_AutoML_2_20220824_183759',
  0.17464672029018402,
  0.6340053990579345,
  0.24383436965802252),
 ('GLM_2_AutoML_2_20220824_183759', 0.0, 0.0, 0.0),
 ('DeepLearning_grid_1_AutoML_2_20220824_183759_model_2', 0.0, 0.0, 0.0)]
```

In the case of Super Learner "StackedEnsemble_**BestOfFamily**_7_AutoML_2_20220824_183759" considering best base models one from each family, out of that it gives more importance to GBM_grid_1 and least to DeepLearning_grid as a features.

```
model.model_performance(test)
```

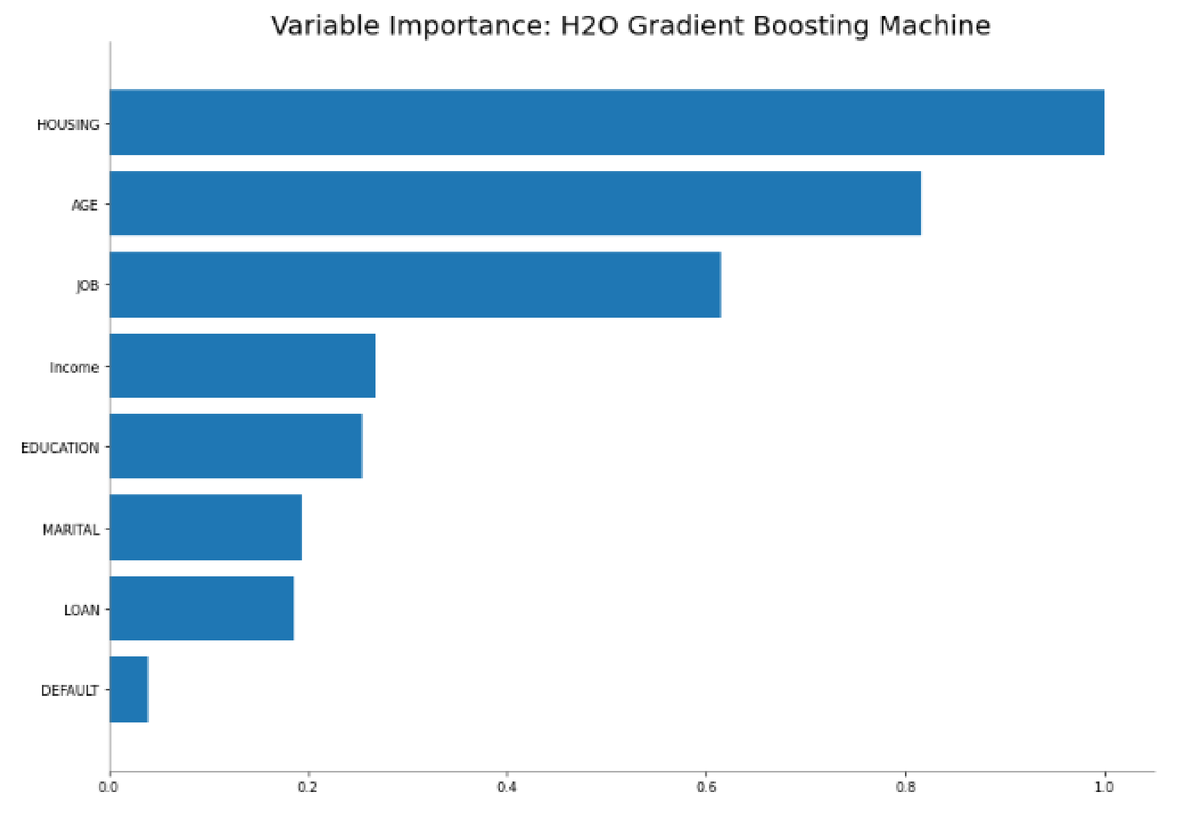
```
ModelMetricsBinomial: gbm
** Reported on test data. **

MSE: 0.10050197232167722
RMSE: 0.3170204604149032
LogLoss: 0.3443474364669525
Mean Per-Class Error: 0.3687871665059467
AUC: 0.6950938119441765
AUCPR: 0.26415888202544396
Gini: 0.3901876238835304
```

In the case of Base models best model is : **GBM**_grid_1_AutoML_2_20220824_183759_model_3, with Accuracy is 87.81% and AUC : 69.51%

Model Evaluation

Results | Interpretation | Insights



Variable importance given by best base model (GBM_grid_1_AutoML_2_20220824_183759_model_3), its treating Housing, Age and Job are playing most important roles in deciding Loan Approval

Model Evaluation

Results | Interpretation | Insights

Explain a single row prediction

The h2o.explain_row() function provides model explanations for a single row of test data. You can evaluate row-level behavior by specifying the row_index

```
print(test[25,:])
print(predictions[25,:])
```

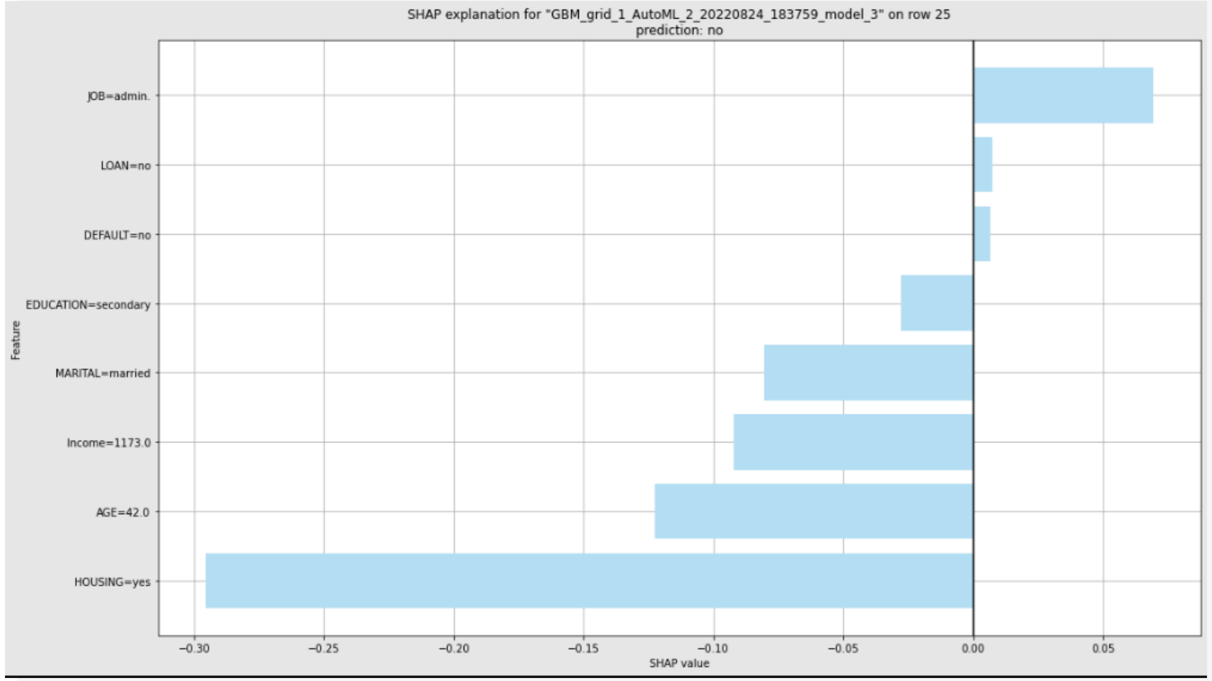
AGE	JOB	MARITAL	EDUCATION	DEFAULT	HOUSING	LOAN	LOAN_STATUS	Income
42	admin.	married	secondary	no	yes	no	no	1173

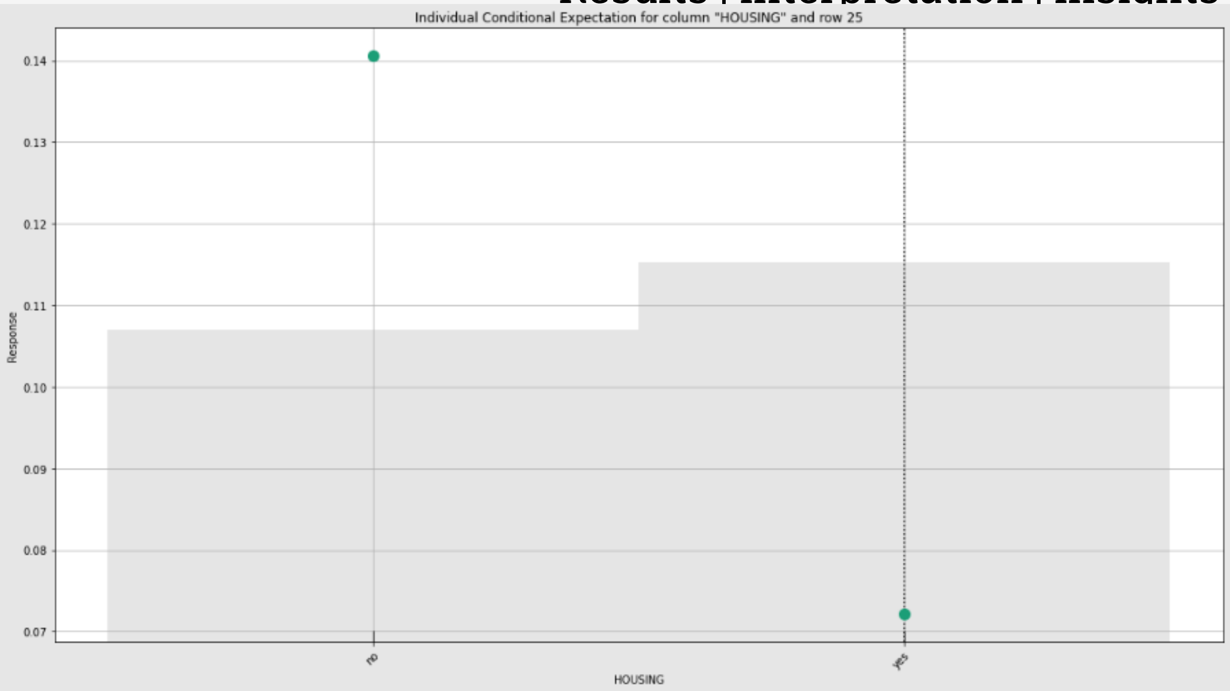
predict	no	yes
no	0.927775	0.0722251

As per the dataset Actual value of row index 25th is “No” and Best base model is giving a probability of 92.77 % of rejection.

Model Evaluation

Results | Interpretation | Insights



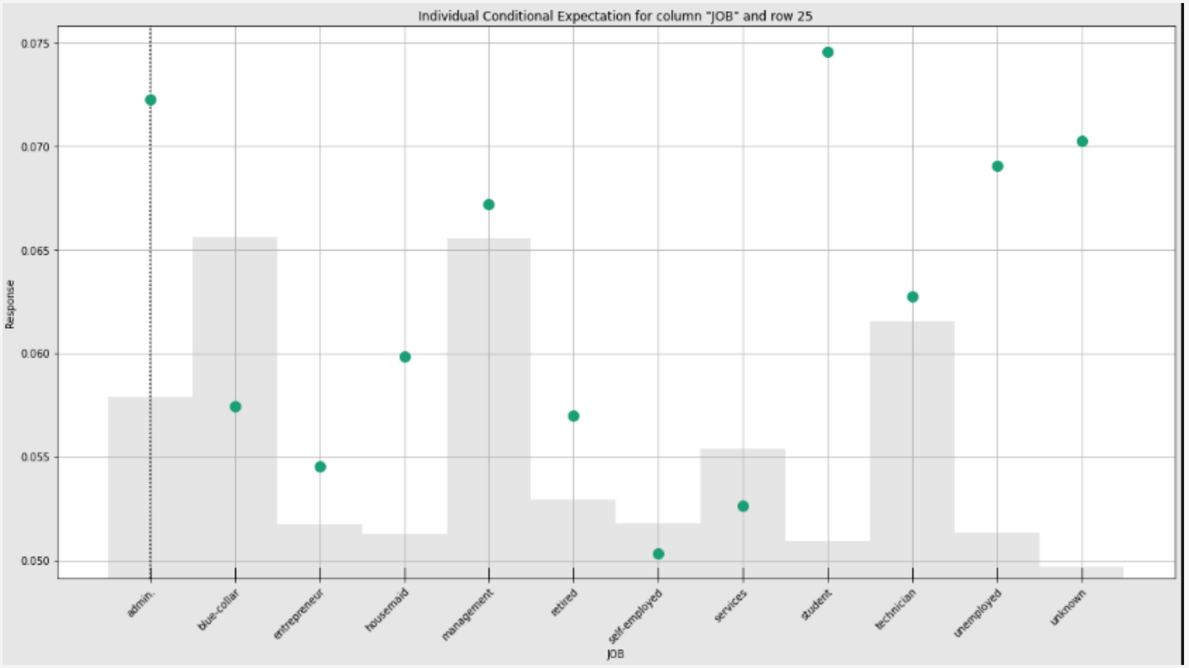


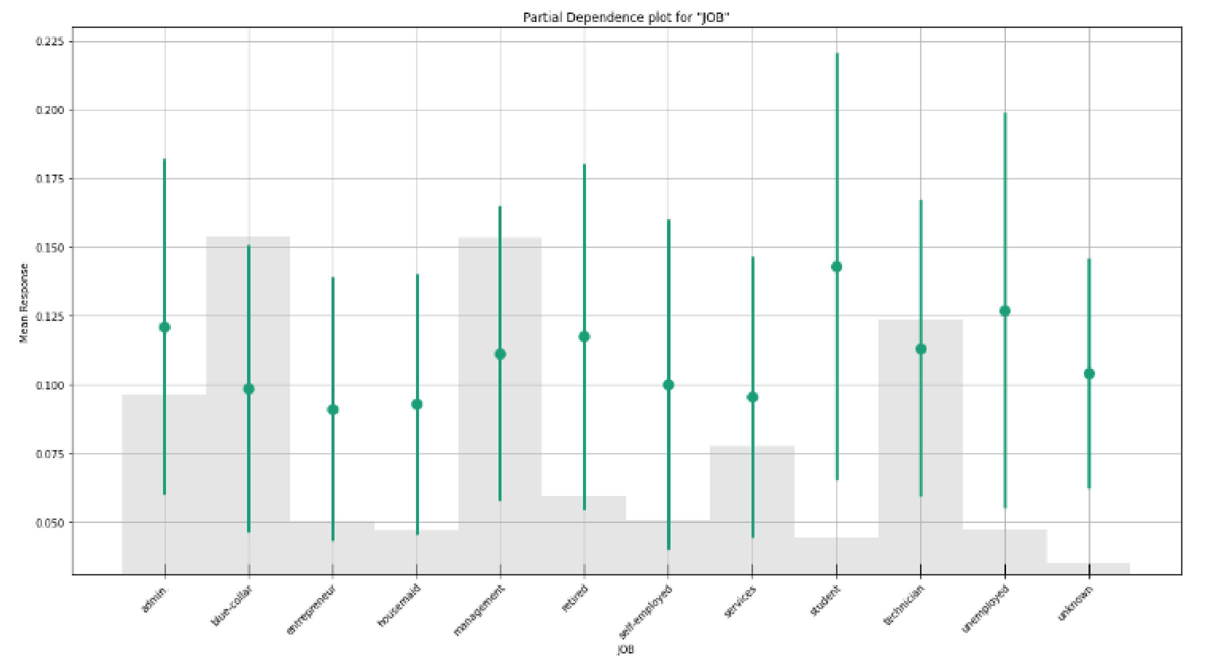
SHAP explanation shows contribution of features for a given instance. The sum of the feature contributions and the bias term is equal to the raw prediction of the model

Individual conditional expectations (ICE) plot gives a graphical depiction of the marginal effect of a variable on the response for a given row.

Model Evaluation

Results | Interpretation | Insights



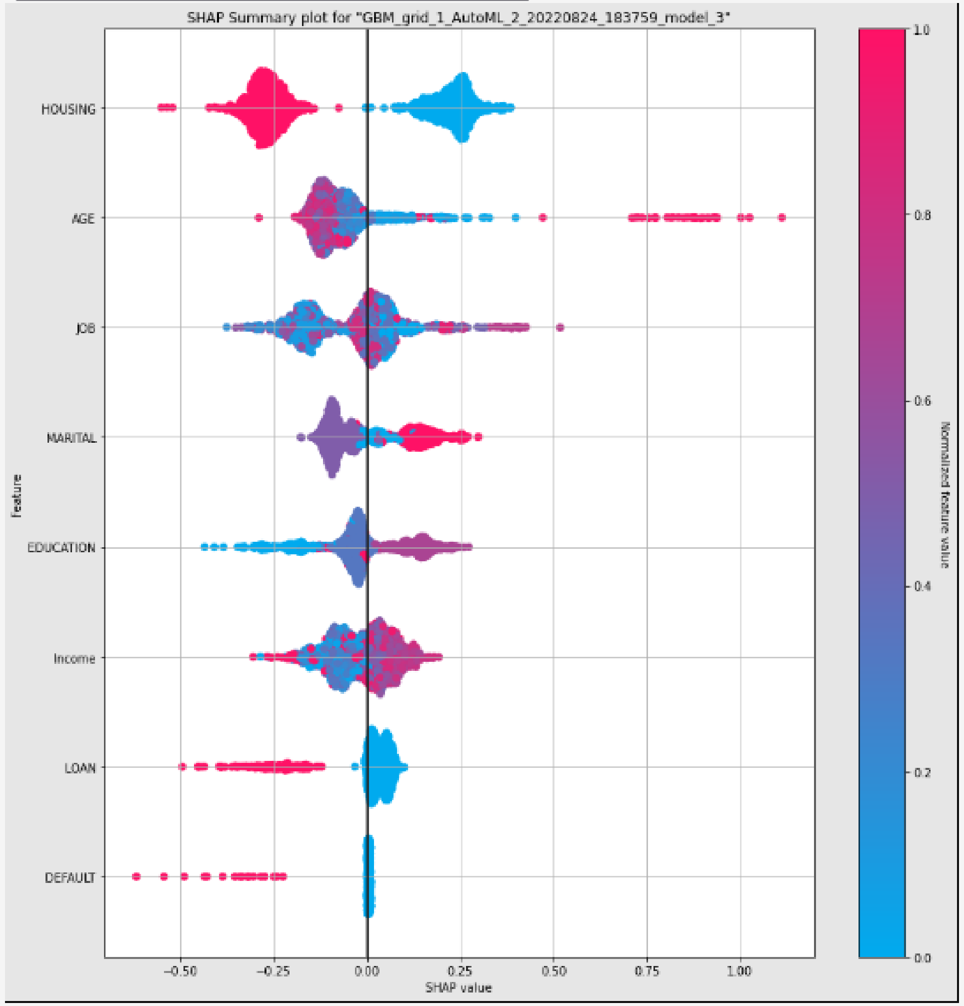


Individual conditional expectations (ICE) plot gives a graphical depiction of the marginal effect of a variable on the response for a given row.

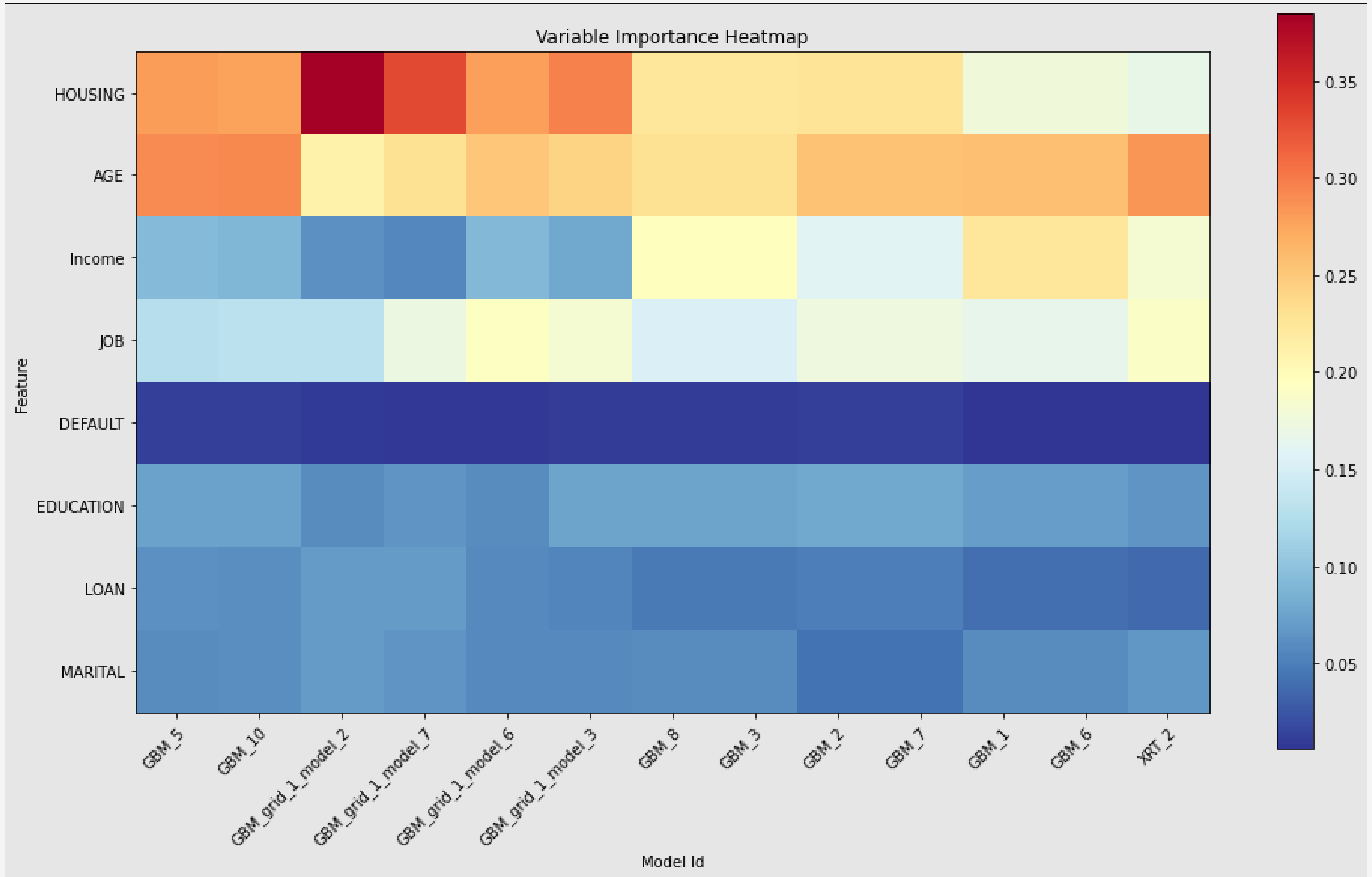
Partial dependence plot (PDP) gives a graphical depiction of the marginal effect of a variable on the response. PDP assumes independence between the feature for which is the PDP computed and the rest.

REVA UNIVERSITY
Bengaluru, India

Model Evaluation



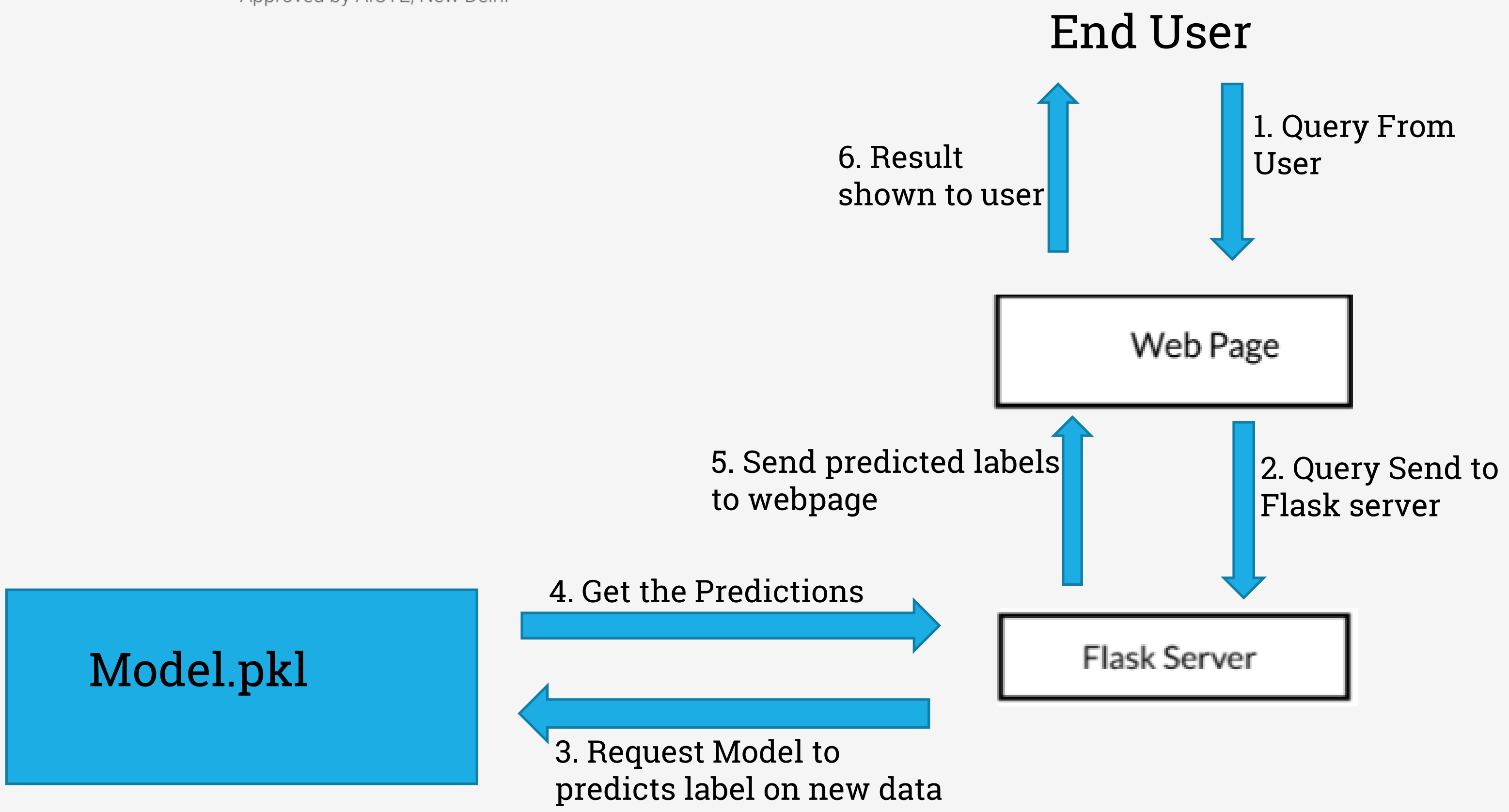
Positive SHAP value means positive impact on prediction, leading the model to predict 1(e.g. Loan approved). Negative SHAP value means negative impact.



Variable Importance Heatmap : Variable importance heatmap shows variable importance across multiple models.

Model Deployment

Demonstration



We are planning to deploy the model on the flask server with the help of pickel file of the saved model which can predict the category of Employee.

Results and Insights

Key Findings | Suggestions

On the present data of credit risk data set, with help of AutoML there are 38 statistical has been developed including Super learner as well the base models. With the best base models which was GBM grid we found the accuracy of 87.81% and AUC as 69.50%. In case of Best super learner which is the “StackedEnsemble_BestOfFamily” has the test Accuracy 87.86% and AUC : 69.94%.

In this research, it has been shown that with the use of Automate Machine learning technique in the combination of different explanations e.g. SHAP, PDP and ICE. Different complex, as well as base models, can be developed in a short time with minimal knowledge of programming and compared with different metrics. Researchers could save a lot of time in developing, training, or tuning the different machine learning models, they could spend that time on data collection and understanding it. On rejecting any loan application end user can explain the reason or features behind that so the customer can also be satisfied with the explanation.

Conclusion and Future Work

Proposed solutions | Scope for future work

Hence in this project we've developed many high end complex machine learning statistical models in very short time with development, training and their hypertunning. With their metrics user can compare them and select the best model to suit to their requirement. In this AutoML also explain the role of each input variable in the prediction of different high end complex models.

With the help of SHAP, ICE and PDP user can explain each individual prediction by any base as well Super learner machine learning models. So the financial/ Banking institution can give clear explanation to their customers why their loan got rejected or approved.

In this project we have used the sample dataset from the internet as the customer information is very confidential property in any financial institution. In case if we get some real data in future we could test this approach on that. It can be very useful in assessing the credit risk while approving any loan application as well giving the better interpretation of any individual prediction also.

References

Bibliography | Webliography

❖BCBS. (2003). Basel II: The New Basel Capital Accord, Basel Committee on Banking Supervision. *https://www.bis.org/bcbs/bcbscp3.htm* , 226.

❖CIRCULARS, R. (2008). *RBI CIRCULARS*. Retrieved from Reserve Bank Of India: https://rbi.org.in/Scripts/BS_CircularIndexDisplay.aspx?Id=4682

❖Christian Bluhm, L. O. (2010). *Introduction to Credit Risk Modeling*. New York: Taylor & Francis Group.

❖Tomasz R. Bielecki, M. R. (2013). *Credit Risk: Modeling, Valuation and Hedging*. New York: Springer Finance.

❖Omomehin, V. (2021). *Credit Risk: Implementing Structural Models*. Cork: AIMS.

❖Pandey, P. (2019). A Deep dive into H2O's AutoML. *https://towardsdatascience.com/a-deep-dive-into-h2os-automl-4b1fe51d3f3e* .

❖Chatterjee, S. (2022). Modelling Credit Risk | Bank of England. *https://www.bankofengland.co.uk/ccbs/modelling-credit-risk* .

❖BIS. (2009). Basel II: Revised international capital framework, Basel Committee on Banking Supervision. *https://www.bis.org/publ/bcbsca.htm* .

❖Narasimham, M. (n.d.). Narasimham Committee. https://en.wikipedia.org/wiki/Narasimham_Committee .

Annexure

Additional Information | Plagiarism score

Similarity Index Report

- Software Used : **Turnitin**
- Date of Report Generation : **26th - Aug-2022**
- Similarity Index in % : **11%**
- Total word count: **9,033**
- Name of the Guide : **Ravi Shukla**

Investigating Super learner for Credit Risk Modeling in Mortgage Scenario

by Lalit Aggarwal

Submission date: 26-Aug-2022 10:37AM (UTC+0530)
Submission ID: 1887290814
File name: arner_for_Credit_Risk_Modeling_in_Mortgage_Scenario_-_Lalit.docx (845.42K)
Word count: 9033
Character count: 49485

Investigating Super learner for Credit Risk Modeling in Mortgage Scenario

ORIGINALITY REPORT

11%	7%	1%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

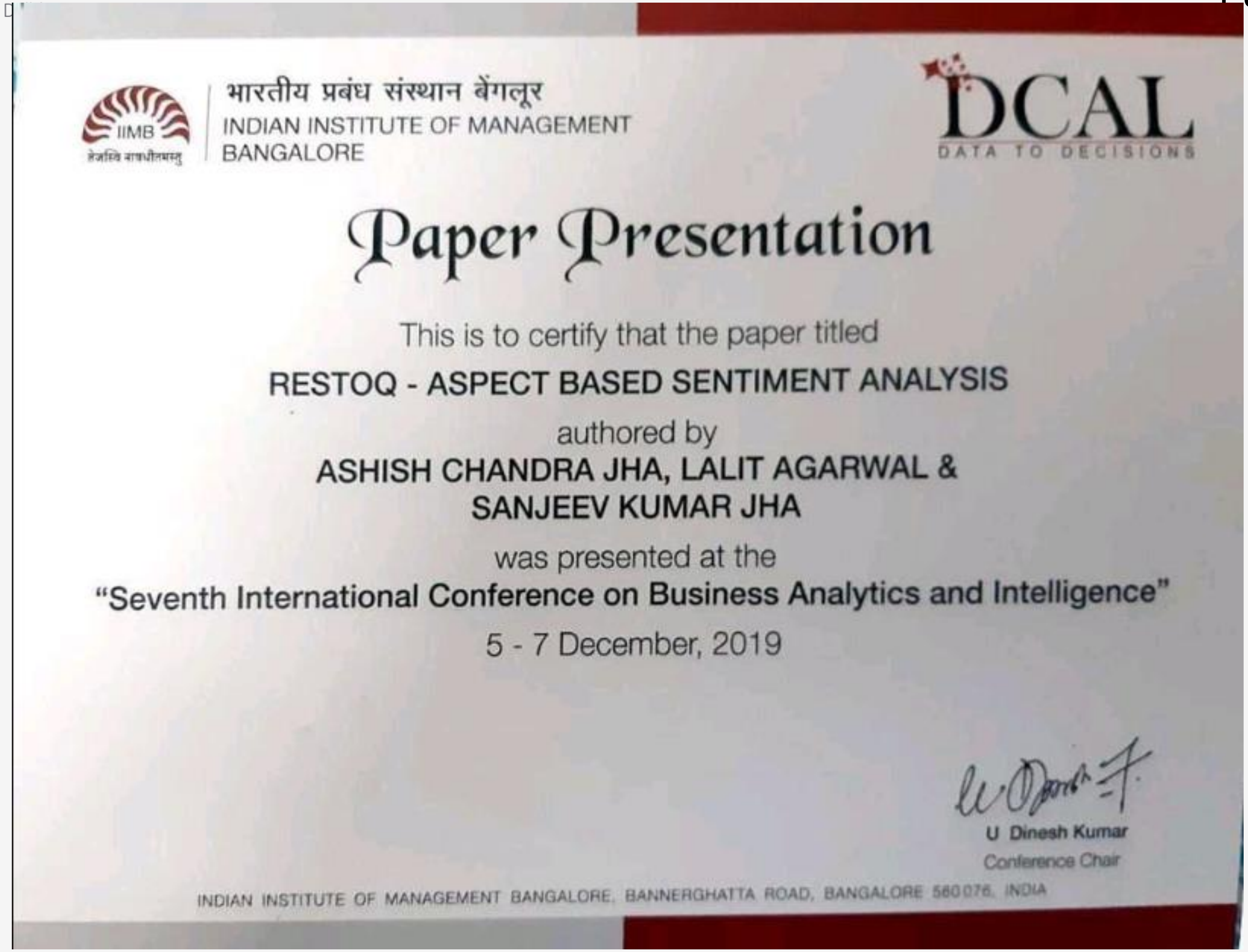
PRIMARY SOURCES

REVA
UNIVERSITY
Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New D

Annexure

Publications | Conferences


भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

DCAL
DATA TO DECISIONS

Paper Presentation

This is to certify that the paper titled

RESTOQ - ASPECT BASED SENTIMENT ANALYSIS

authored by

ASHISH CHANDRA JHA, LALIT AGARWAL &
SANJEEV KUMAR JHA

was presented at the

"Seventh International Conference on Business Analytics and Intelligence"

5 - 7 December, 2019

U Dinesh Kumar
Conference Chair

INDIAN INSTITUTE OF MANAGEMENT BANGALORE, BANNERGHATTA ROAD, BANGALORE 560076, INDIA

REVA UNIVERSITY
Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Special thanks to Ravi Shukla, Dr. J. B. Simha and Dr. Shinu Abhi for their guidance and all the support on this initiative !!!

Thank you!