LAB MID CAT

Name : Yashika kudesia   Registration No: 23MEC011

DATA ROBOT AN AI TOOL FOR ML

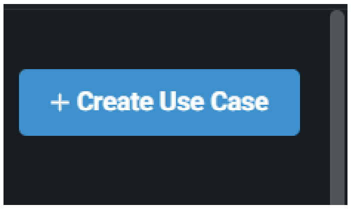
Q:- WHAT IS DATA ROBOT ?

Ans :- DataRobot is an AI platform used to build the predictive models enterprises depend on to accelerate their growth. The platform draws from a number of open-source machine learning R and Python-based libraries, including scikit-learn, H2O, TensorFlow, Vowpal Wabbit, Spark ML, and XGBoost. But with DataRobot's simple, drag-and-drop web-based interface, building and deploying sophisticated predictive models is a breeze…even for business analysts with little-to-no knowledge of machine learning or programming. By automating the selection of the ideal features, algorithms, and parameter values for building each model, the software supports best practices for new users. Meanwhile, the platform remains both flexible and extensible.

Q:- HOW IT WORKS ?

Ans :- Here are the following steps that involve in

Step 1 : Get started with Creating the project



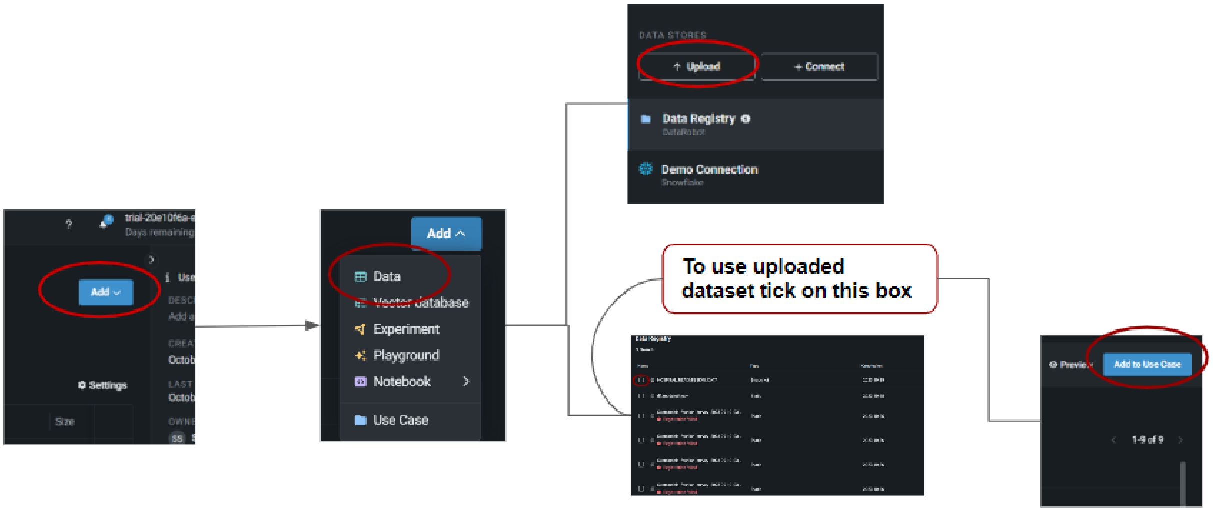
Step 2 : Prepare the Dataset

To prep the data using DataRobot we can import a local dataset or can connect to an external data source.
To complete the quickstart, you first log in to DataRobot Data Prep.
Once you log in, complete these steps:

- ➢ Click on ADD
- ➢ Then Data
- ➢ Upload(can publish your dataset) or Use uploaded dataset.

The hospital readmission dataset for analysis

Can upload our dataset

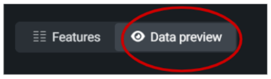


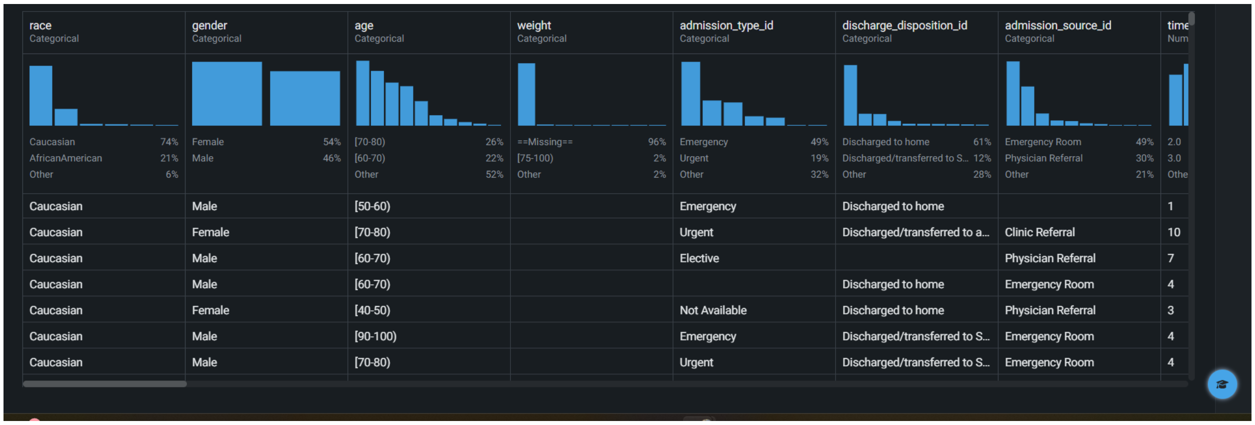
To use uploaded dataset tick on this box

Else can work with already uploaded dataset

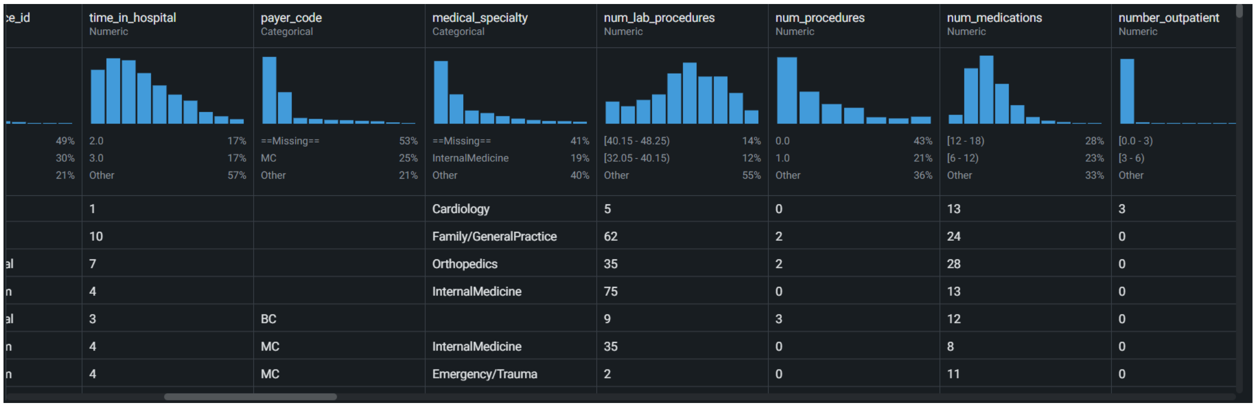
Step 3 : Data visualization

To analyze the dataset, can directly click the dataset.
There can check the histogram and detailed values of the data

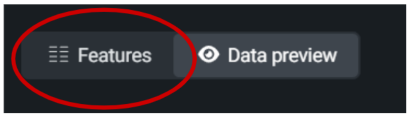


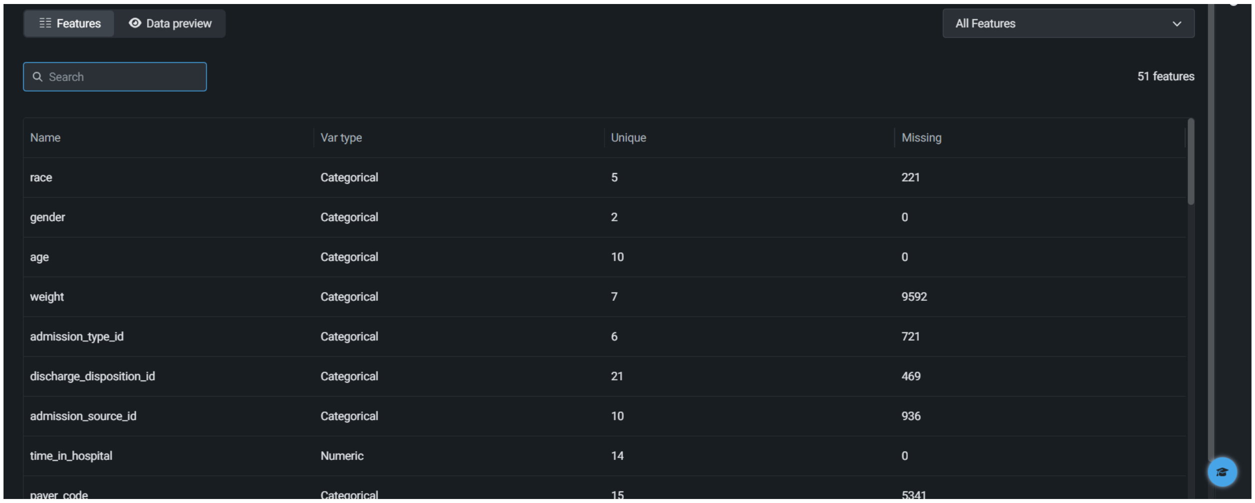


race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time
Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Num
Caucasian 74% AfricanAmerican 21% Other 6%	Female 54% Male 46%	[70-80) 26% [60-70) 22% Other 52%	==Missing== 96% [75-100) 2% Other 2%	Emergency 49% Urgent 19% Other 32%	Discharged to home 61% Discharged/transferred to S... 12% Other 28%	Emergency Room 49% Physician Referral 30% Other 21%	2.0 3.0 Othe
Caucasian	Male	[50-60)		Emergency	Discharged to home		1
Caucasian	Female	[70-80)		Urgent	Discharged/transferred to a...	Clinic Referral	10
Caucasian	Male	[60-70)		Elective		Physician Referral	7
Caucasian	Male	[60-70)			Discharged to home	Emergency Room	4
Caucasian	Female	[40-50)		Not Available	Discharged to home	Physician Referral	3
Caucasian	Male	[90-100)		Emergency	Discharged/transferred to S...	Emergency Room	4
Caucasian	Male	[70-80)		Urgent	Discharged/transferred to S...	Emergency Room	4



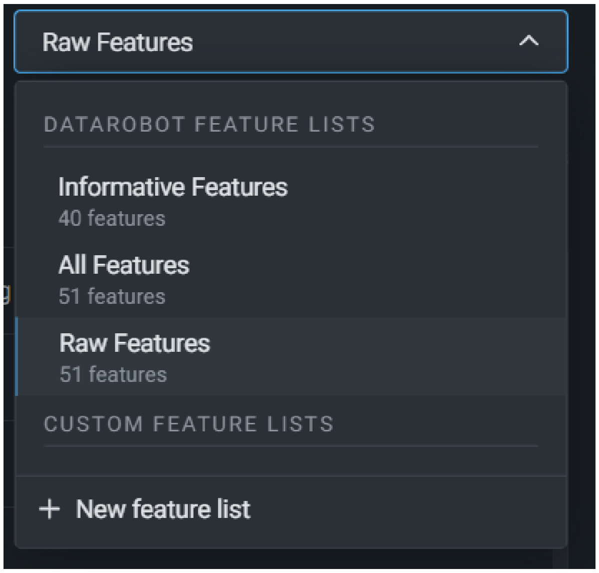
ce_id	time_in_hospital	payer_code	medical_specialty	num_lab_procedures	num_procedures	num_medications	number_outpatient
	Numeric	Categorical	Categorical	Numeric	Numeric	Numeric	Numeric
49% 30% 21%	2.0 17% 3.0 17% Other 57%	==Missing== 53% MC 25% Other 21%	==Missing== 41% InternalMedicine 19% Other 40%	[40.15 - 48.25) 14% [32.05 - 40.15) 12% Other 55%	0.0 43% 1.0 21% Other 36%	[12 - 18) 28% [6 - 12) 23% Other 33%	[0.0 - 3) [3 - 6) Other
	1		Cardiology	5	0	13	3
	10		Family/GeneralPractice	62	2	24	0
al	7		Orthopedics	35	2	28	0
n	4		InternalMedicine	75	0	13	0
al	3	BC		9	3	12	0
n	4	MC	InternalMedicine	35	0	8	0
n	4	MC	Emergency/Trauma	2	0	11	0





Name	Var type	Unique	Missing
race	Categorical	5	221
gender	Categorical	2	0
age	Categorical	10	0
weight	Categorical	7	9592
admission_type_id	Categorical	6	721
discharge_disposition_id	Categorical	21	469
admission_source_id	Categorical	10	936
time_in_hospital	Numeric	14	0
payer_code	Categorical	15	5341

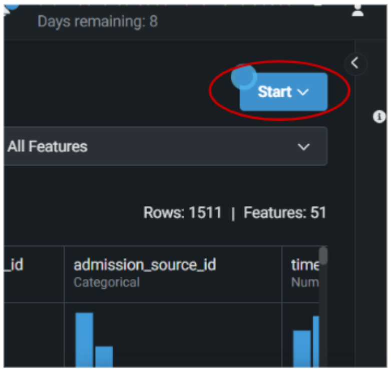
From here we can check different type of feature in our dataset



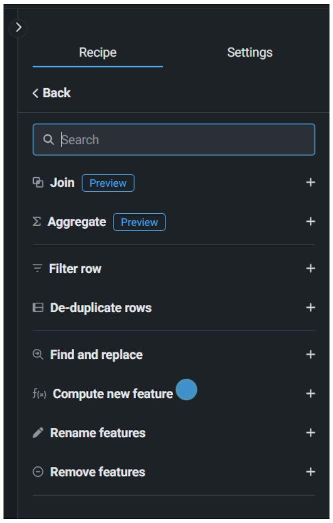
Step 4 : Modeling

When we click on start we will get the two option

- Modeling
- Wrangling

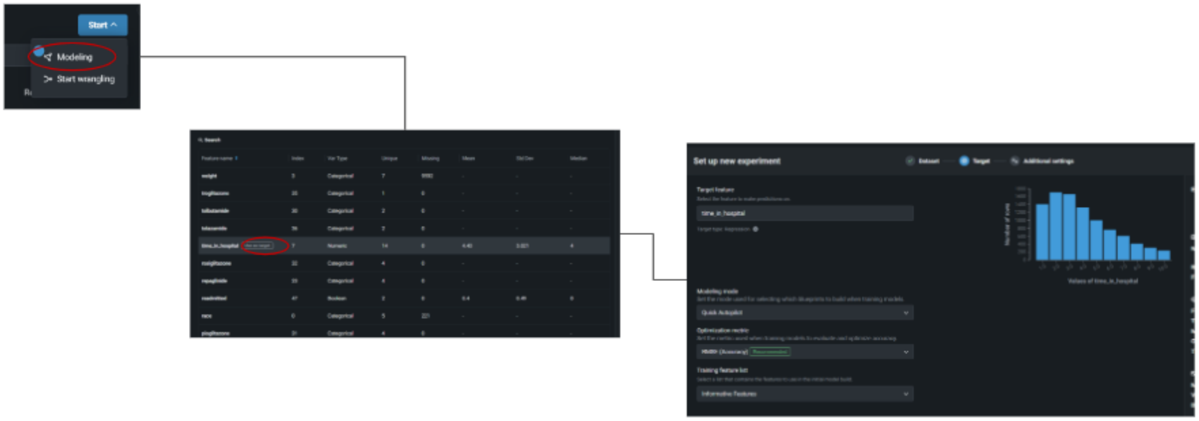


On clicking wrangling : Start wrangling to pull a random sample of data from the data source and begin transformation operations.
Click Add operation : to build a wrangling "recipe." Each new operation updates the live sample to reflect the transformation. Note that if you wrangle your training dataset, you will want to apply the same operations to your scoring dataset to ensure you have the same columns.



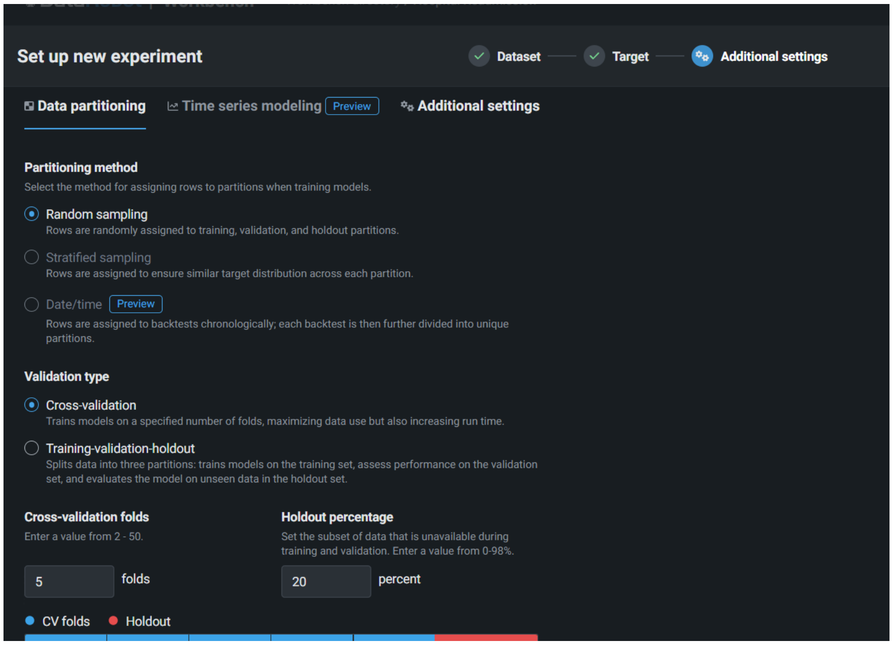
On clicking Modeling : we will get the following function

- ➢ **Click on modeling**
- ➢ **Set the target**
- ➢ **Change the mode of model(if needed)**



- ➢ **Click on additional settings**
- ➢ **In data partitioning**

- ○ **Select the partitioning method**
- ○ **Select the validation type according to your need**



Set up new experiment

✓ Dataset — ✓ Target — Additional settings

Data partitioning | Time series modeling Preview | Additional settings

Partitioning method

Select the method for assigning rows to partitions when training models.

● Random sampling
Rows are randomly assigned to training, validation, and holdout partitions.

○ Stratified sampling
Rows are assigned to ensure similar target distribution across each partition.

○ Date/time Preview
Rows are assigned to backtests chronologically; each backtest is then further divided into unique partitions.

Validation type

● Cross-validation
Trains models on a specified number of folds, maximizing data use but also increasing run time.

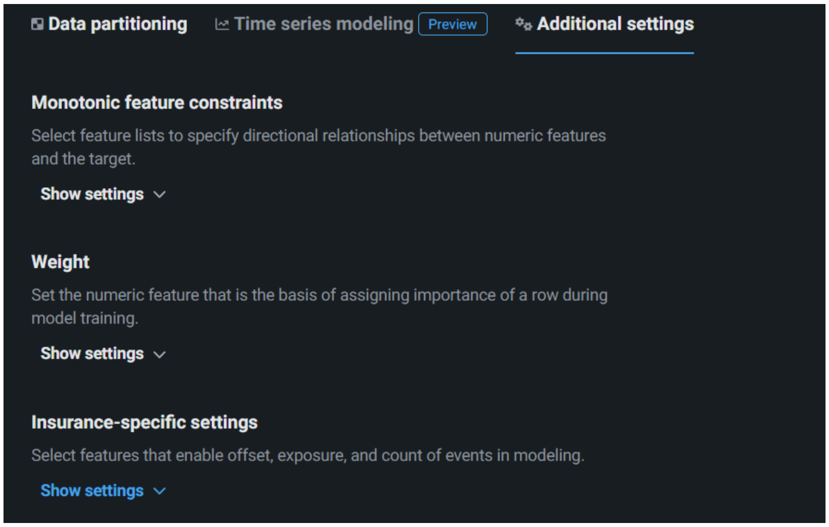
○ Training-validation-holdout
Splits data into three partitions: trains models on the training set, assess performance on the validation set, and evaluates the model on unseen data in the holdout set.

Cross-validation folds
Enter a value from 2 - 50.
5 folds

Holdout percentage
Set the subset of data that is unavailable during training and validation. Enter a value from 0-98%.
20 percent

● CV folds ● Holdout

- ➢ **In additional settings**



Data partitioning | Time series modeling Preview | Additional settings

Monotonic feature constraints

Select feature lists to specify directional relationships between numeric features and the target.

Show settings ⌄

Weight

Set the numeric feature that is the basis of assigning importance of a row during model training.

Show settings ⌄

Insurance-specific settings

Select features that enable offset, exposure, and count of events in modeling.

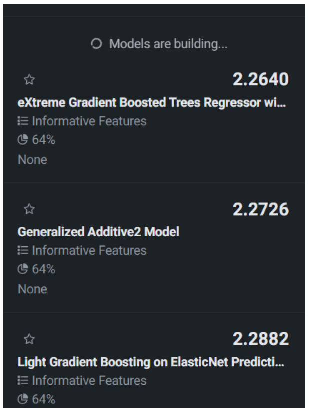
Show settings ⌄

- ➢ **Click start modeling**

Step 5 : Model and Outputs

Once modeling starts, Workbench begins to construct a model Leaderboard. Ultimately, DataRobot will select and retrain the most accurate model and mark it as prepared for deployment. Since the process takes some time, click on any completed model and familiarize yourself with the insights available for model evaluation.

➢ Select the model

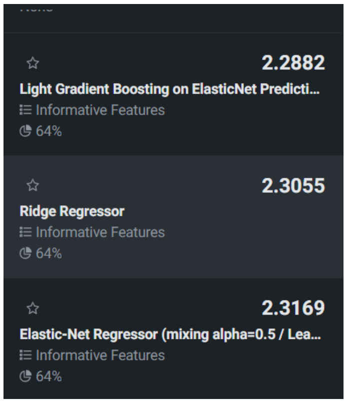


Models are building...

☆ 2.2640
eXtreme Gradient Boosted Trees Regressor wi...
Informative Features
64%
None

☆ 2.2726
Generalized Additive2 Model
Informative Features
64%
None

☆ 2.2882
Light Gradient Boosting on ElasticNet Predicti...
Informative Features
64%



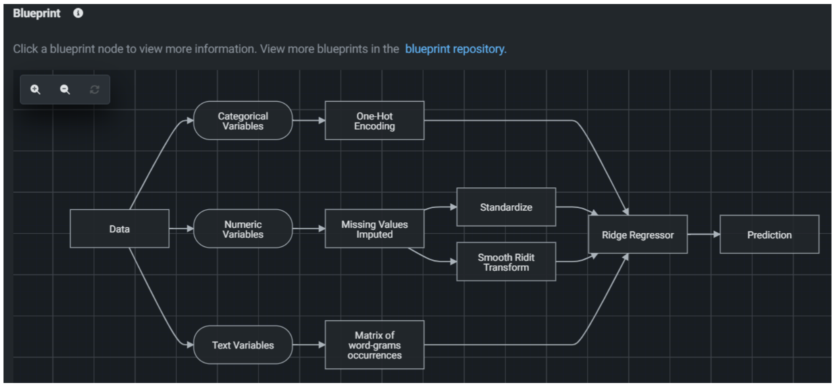
☆ 2.2882
Light Gradient Boosting on ElasticNet Predicti...
Informative Features
64%

☆ 2.3055
Ridge Regressor
Informative Features
64%

☆ 2.3169
Elastic-Net Regressor (mixing alpha=0.5 / Lea...
Informative Features
64%

➢ After selection the model can see the following feature

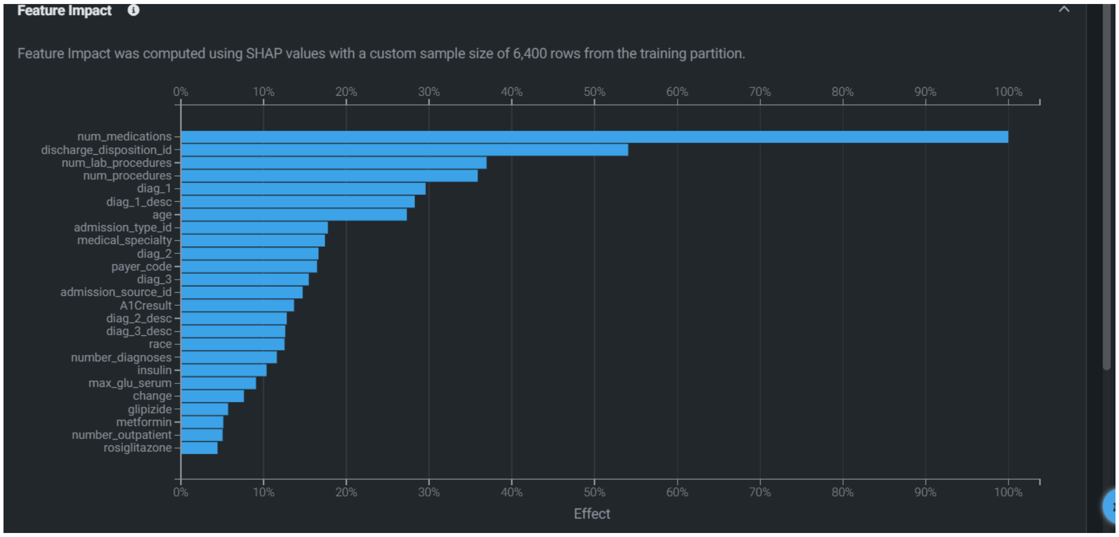
○ Blueprint of the model



Blueprint

Click a blueprint node to view more information. View more blueprints in the blueprint repository.

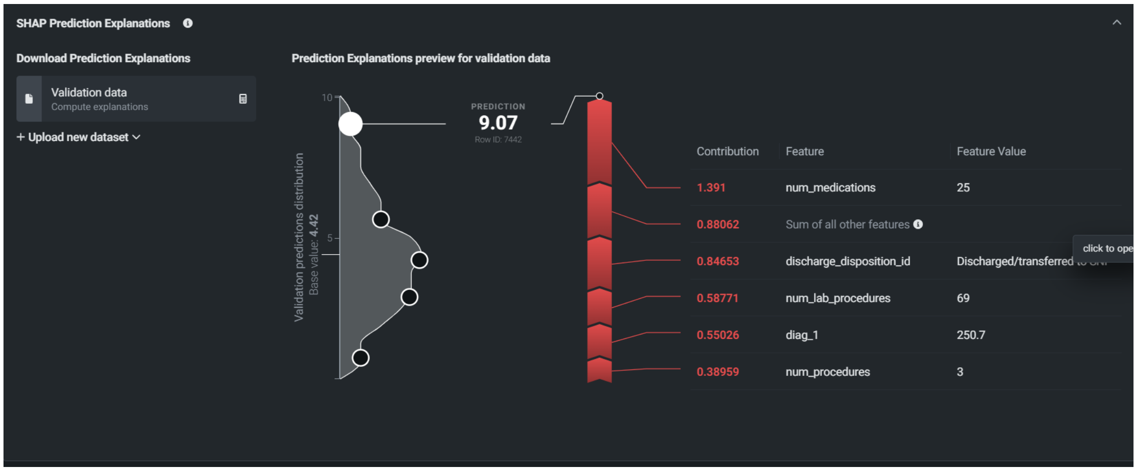
○ Feature Impact



Feature Impact

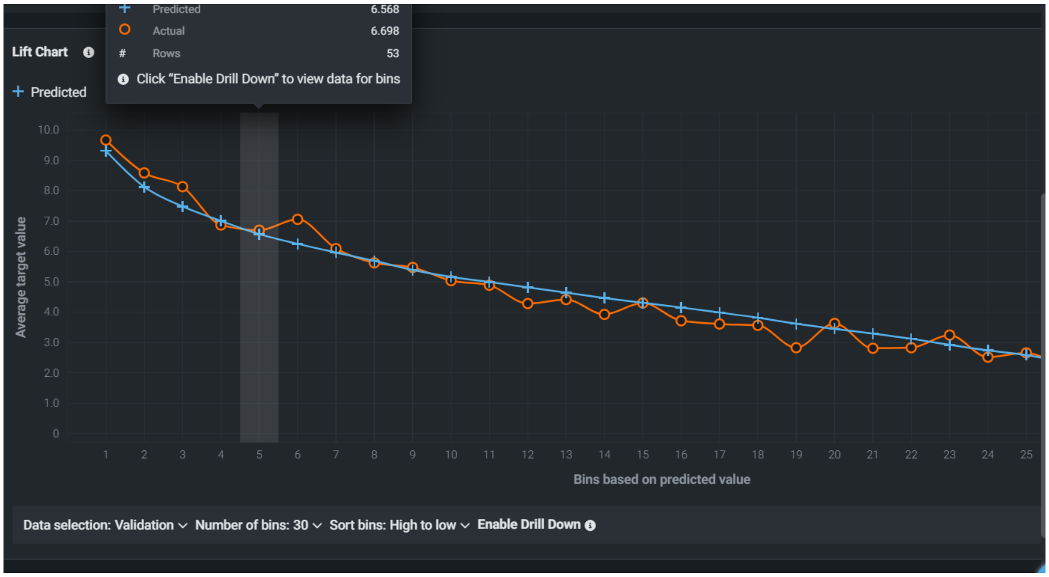
Feature Impact was computed using SHAP values with a custom sample size of 6,400 rows from the training partition.

○ Prediction Explanation


SHAP Prediction Explanations
Download Prediction Explanations
Validation data
Compute explanations
+ Upload new dataset
Prediction Explanations preview for validation data
10
5
Validation predictions distribution
Base value: 4.42
PREDICTION
9.07
Row ID: 7442
click to ope

Contribution	Feature	Feature Value
1.391	num_medications	25
0.88062	Sum of all other features	
0.84653	discharge_disposition_id	Discharged/transferred
0.58771	num_lab_procedures	69
0.55026	diag_1	250.7
0.38959	num_procedures	3

○ Lift chart

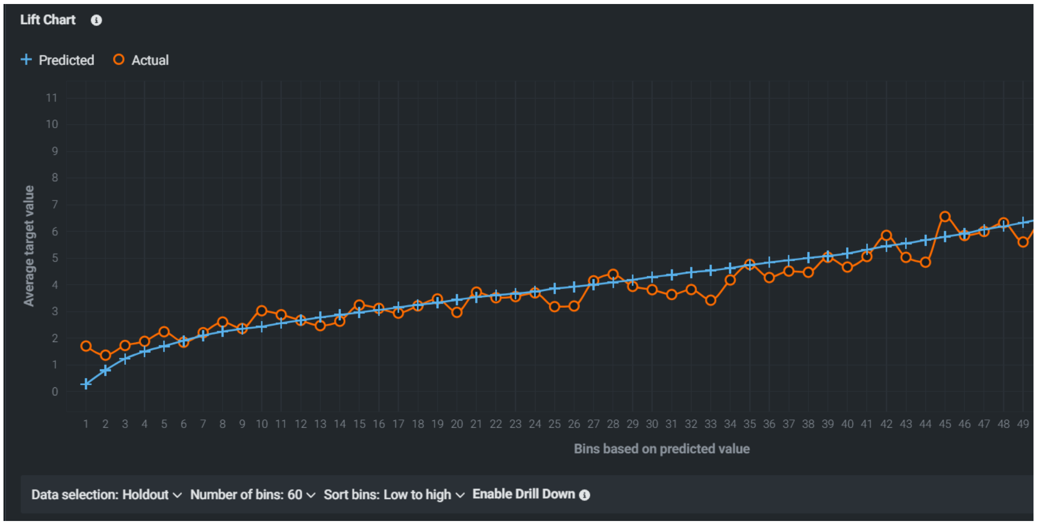


+ Predicted 6.568
O Actual 6.698
# Rows 53

Click "Enable Drill Down" to view data for bins

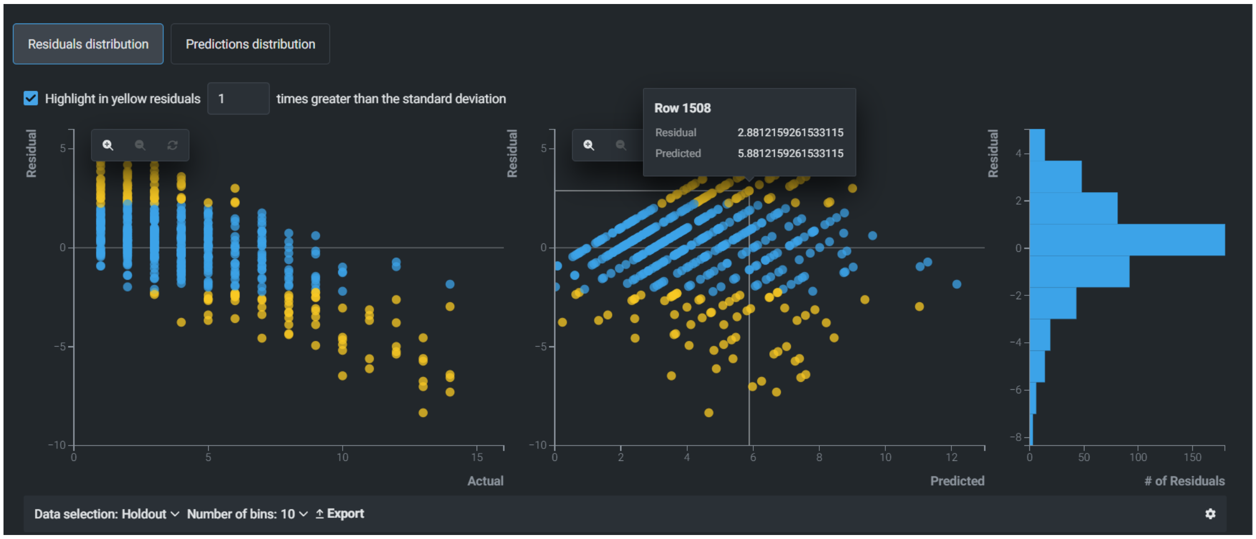
Data selection: Validation ⌄ Number of bins: 30 ⌄ Sort bins: High to low ⌄ Enable Drill Down

The value of the number of bins, data selection and sort bin can be change



Data selection: Holdout ⌄ Number of bins: 60 ⌄ Sort bins: Low to high ⌄ Enable Drill Down

- ### Residuals



☑ Highlight in yellow residuals 1 times greater than the standard deviation

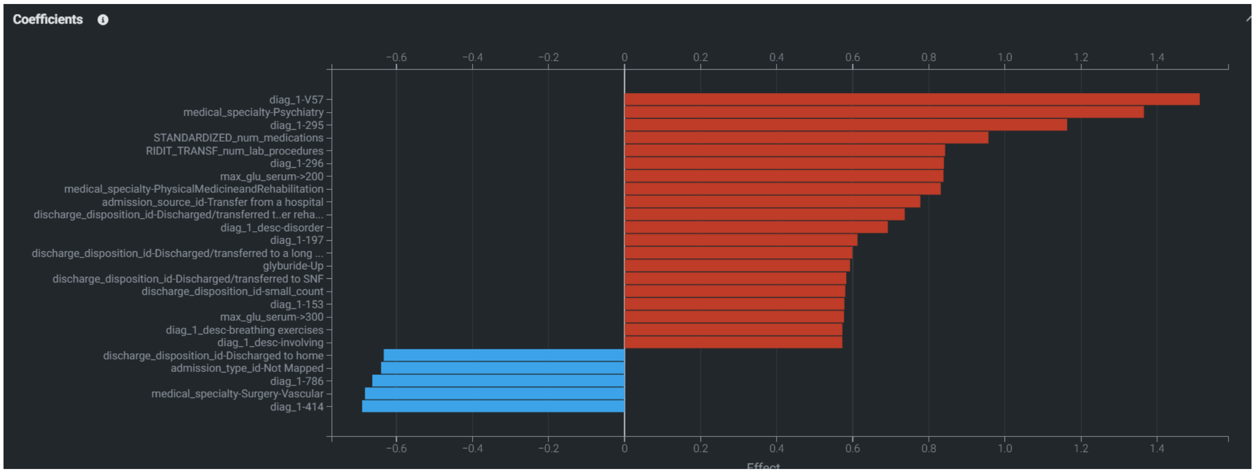
Row 1508

Residual 2.8812159261533115

Predicted 5.8812159261533115

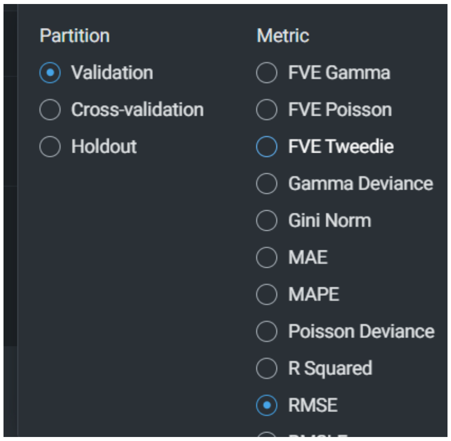
Data selection: Holdout ⌄ Number of bins: 10 ⌄ ⇧ Export

- ### Coefficient


Coefficients
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
diag_1-V57
medical_specialty-Psychiatry
diag_1-295
STANDARDIZED_num_medications
RIDIT_TRANSF_num_lab_procedures
diag_1-296
max_glu_serum->200
medical_specialty-PhysicalMedicineandRehabilitation
admission_source_id-Transfer from a hospital
discharge_disposition_id-Discharged/transferred t..er reha...
diag_1_desc-disorder
diag_1-197
discharge_disposition_id-Discharged/transferred to a long ...
glyburide-Up
discharge_disposition_id-Discharged/transferred to SNF
discharge_disposition_id-small_count
diag_1-153
max_glu_serum->300
diag_1_desc-breathing exercises
diag_1_desc-involving
discharge_disposition_id-Discharged to home
admission_type_id-Not Mapped
diag_1-786
medical_specialty-Surgery-Vascular
diag_1-414
Effect

➢ On the partition (validation, cross validation & holdout)

- ### Can select the matrix according to need



Partition

● Validation

○ Cross-validation

○ Holdout

Metric

○ FVE Gamma

○ FVE Poisson

○ FVE Tweedie

○ Gamma Deviance

○ Gini Norm

○ MAE

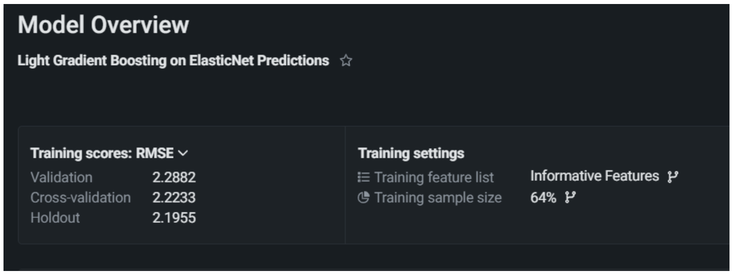
○ MAPE

○ Poisson Deviance

○ R Squared

● RMSE

- ○ **Model overview**



Model Overview

Light Gradient Boosting on ElasticNet Predictions ☆

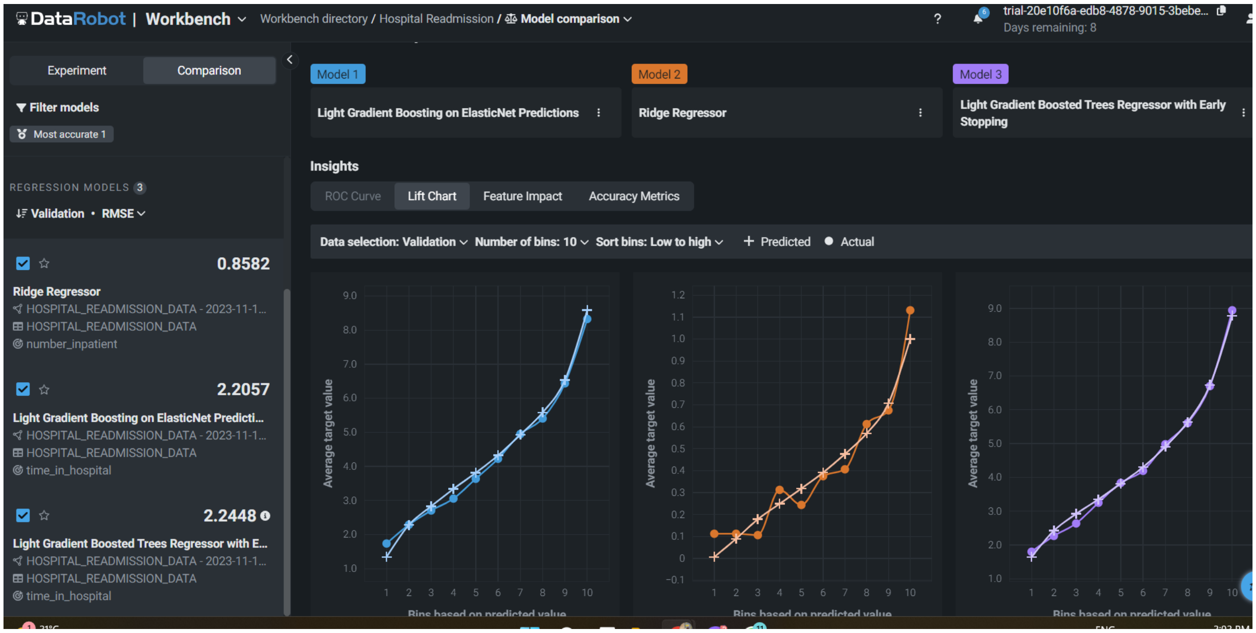
Training scores: RMSE

Validation	2.2882
Cross-validation	2.2233
Holdout	2.1955

Training settings

Training feature list	Informative Features
Training sample size	64%

In this different type of model and their prediction can also be compared



DataRobot | Workbench

Workbench directory / Hospital Readmission / Model comparison

trial-20e10f6a-edb8-4878-9015-3bebe...

Days remaining: 8

Experiment | Comparison

Filter models

Most accurate 1

REGRESSION MODELS 3

Validation • RMSE

☑ ☆ 0.8582

Ridge Regressor

HOSPITAL_READMISSION_DATA - 2023-11-1...

HOSPITAL_READMISSION_DATA

number_inpatient

☑ ☆ 2.2057

Light Gradient Boosting on ElasticNet Predicti...

HOSPITAL_READMISSION_DATA - 2023-11-1...

HOSPITAL_READMISSION_DATA

time_in_hospital

☑ ☆ 2.2448

Light Gradient Boosted Trees Regressor with E...

HOSPITAL_READMISSION_DATA - 2023-11-1...

HOSPITAL_READMISSION_DATA

time_in_hospital

Model 1: Light Gradient Boosting on ElasticNet Predictions

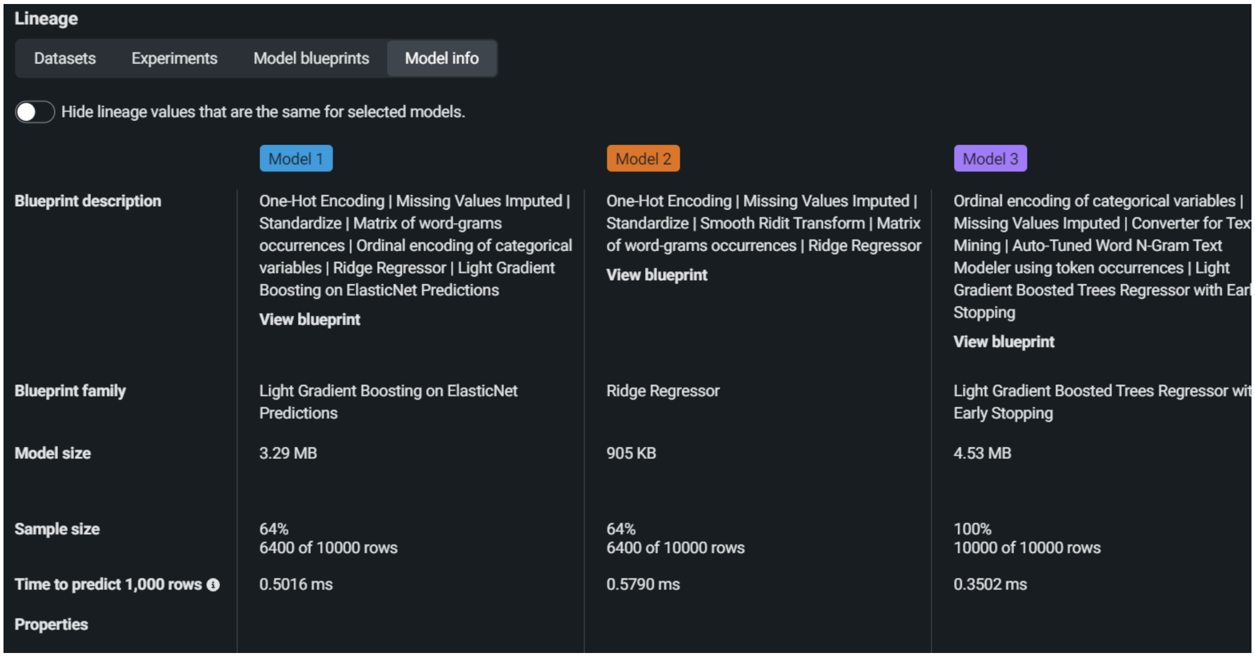
Model 2: Ridge Regressor

Model 3: Light Gradient Boosted Trees Regressor with Early Stopping

Insights

ROC Curve | Lift Chart | Feature Impact | Accuracy Metrics

Data selection: Validation | Number of bins: 10 | Sort bins: Low to high | + Predicted | ● Actual



Lineage

Datasets | Experiments | Model blueprints | Model info

Hide lineage values that are the same for selected models.

	Model 1	Model 2	Model 3
Blueprint description	One-Hot Encoding | Missing Values Imputed | Standardize | Matrix of word-grams occurrences | Ordinal encoding of categorical variables | Ridge Regressor | Light Gradient Boosting on ElasticNet Predictions **View blueprint**	One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Matrix of word-grams occurrences | Ridge Regressor **View blueprint**	Ordinal encoding of categorical variables | Missing Values Imputed | Converter for Tex Mining | Auto-Tuned Word N-Gram Text Modeler using token occurrences | Light Gradient Boosted Trees Regressor with Ear Stopping **View blueprint**
Blueprint family	Light Gradient Boosting on ElasticNet Predictions	Ridge Regressor	Light Gradient Boosted Trees Regressor wi Early Stopping
Model size	3.29 MB	905 KB	4.53 MB
Sample size	64% 6400 of 10000 rows	64% 6400 of 10000 rows	100% 10000 of 10000 rows
Time to predict 1,000 rows	0.5016 ms	0.5790 ms	0.3502 ms
Properties			