Predict term deposit using Dataiku

Submitted:
January 18th, 2024

By:
Shruti Patil: 102138
Seyedmostafa Musavi: 102153
Berlin, Germany

Reviewer:
Prof. Patrick Erdelt

Berliner Hochschule für Technik (BHT)

ABSTRACT

The information relates to a Portuguese banking institution's direct marketing initiatives. Phone calls served as the basis for the marketing initiatives. To determine whether or not the product (bank term deposit) would be subscribed, it was frequently necessary to make multiple contacts with the same consumer. Anticipating whether the client will sign up for a term deposit is the classification goal.

This project aims to develop a predictive model that can accurately identify clients who are likely to subscribe to the product (bank term deposit). To achieve this goal, a dataset containing information about clients will be analysed using Dataiku.

Dataiku is a powerful platform that allows teams to collaborate on data science projects. It is accessible to users with different levels of technical experience and covers everything from creating and deploying Machine Learning models to preparing and exploring data.

By using Dataiku's powerful data mining and machine learning capabilities, we can extract valuable insights from the dataset and build a predictive model that can accurately identify whether the client will sign up for a term deposit. The model will be trained on historical data to learn the patterns and characteristics of clients who signed up for a term deposit. Once the model is trained, it can be used to predict who might sign up for a term deposit based on the characteristics.

Data Understanding & Key Findings from the data

Data:

The given dataset is Portuguese Bank Telemarketing campaign. There are 45 211 observations of 17 features in our data, 10 of which are categorical features and 7 of which are numerical features. The column "y" is the target variable. Picture 1, shows a few samples of the studied data set. Table 1 lists all of the feature names along with their types and values that have been taken.

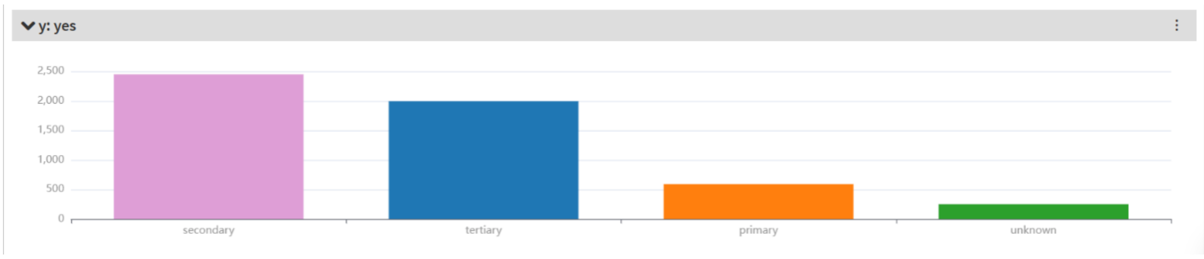
1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
2	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
3	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
4	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
5	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
6	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
7	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
8	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
9	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
10	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
11	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
12	41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
13	29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
14	53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
15	58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
16	57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
17	51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
18	45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
19	57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
20	60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
21	33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
22	28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
23	56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no
24	32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no
25	25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no
26	40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no

Description of features:

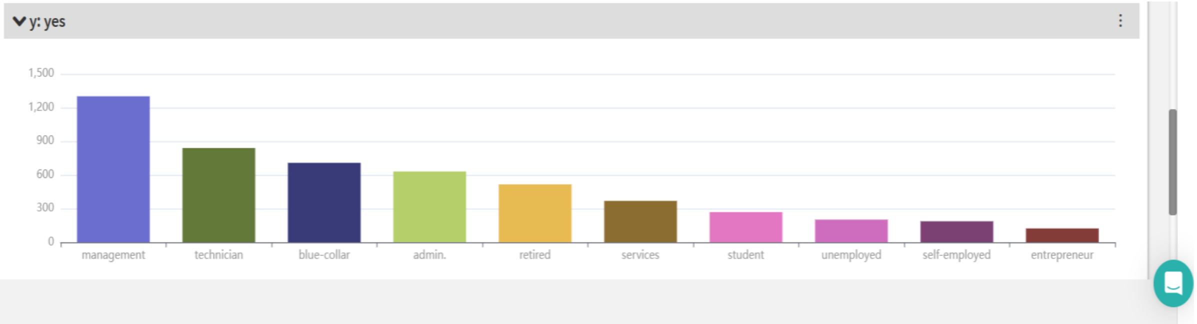
Attributes	Kind	Attribute illustration, description	Values of attributes
age	numeric	age of client	values between 18 and 95
job	categorical	type of job	’management’, ’technician’, ’entrepreneur’, ’blue-collar’, ’unknown’, ’retired’,’admin.’, ’services’, ’self-employed’, ’unemployed’, ’housemaid’, ’student’
marital	categorical	marital status, note: ’divorced’ means divorced or widowed	’divorced’, ’married’, ’single’
education	categorical	degree of education	primary’, ’secondary’, ’tertiary’, ’unknown’
default	binary	has credit in default?	’no’, ’yes’
balance	numeric	account balance	values between -8019 and 102127
housing	binary	has housing loan?	’no’, ’yes’
loan	binary	has personal loan?	’no’, ’yes’
contact	categorical	contact communication type	cellular’, ’telephone, ’unknown’
day	numeric	day in month	Values between 1 and 31
month	categorical	last contact month of year	’Jan’, ’Feb’, ’Mar’, ’Apr’, ’May’, ’Jun’, ’Jul’, ’Aug’, ’Sep’, ’Oct’, ’Nov’, ’Dec’
duration	numeric	last contact duration, in seconds	values between 0 and 4918
campaign	numeric	number of contacts performed during this campaign and for this client (included last contact)	values between 1 and 63
p-days	numeric	number of days that passed by after the client was last contacted from a previous campaign, note: 999 means client was not previously contacted	values between -1 and 871
previous	numeric	umber of contacts performed before this campaign and for this client	values between 0 and 275
p-outcome	categorical	outcome of the previous marketing campaign	’failure’, ’other’, ’success’, ’unknown’
y	binary	has the client subscribed a term deposit?	’no’, ’yes’

2. Key Findings:

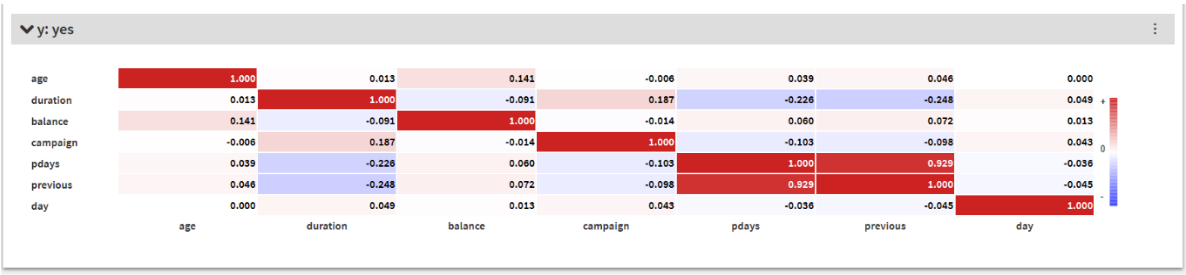
1) From overall data, it is found that the given graph shows that the clients who had secondary and tertiary education, they subscribed to a term deposit.



2) In addition to that, the clients who are management and technician are subscribing more than clients who are entrepreneur and self-employed.



3) In the given Correlation matrix splitted by 'y' with yes, x, we can say that there are 2 columns which are highly correlated with each other(pdays, previous).



3. DATA PREPARATION:

1) DATA CLEAN:

For data cleaning we check if there are null values or duplicates and missing values. We have checked everyone with the Python and Dataiku also.

- **NULL VALUES:**

To find null_values we have used this method df.isnull().sum() and as you can see in picture with the info() method we can see there are no null or missing values.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
```

- **DUPLICATES:**

To check duplicates, we have used duplicates method in python and as it shown in the following picture there are duplicates.

```
In [16]: df.duplicated()

Out[16]: 0        False
         1        False
         2        False
         3        False
         4        False
                  ...
         45206    False
         45207    False
         45208    False
         45209    False
         45210    False
         Length: 45211, dtype: bool

In [18]: df.duplicated().sum()

Out[18]: 0
```

2) LABEL ENCODER:

In data preprocessing phase we have decided to use "Label encoding" method between "Label Encoding" and "OneHotEncoding" for changing categorical columns to numerical. In the following screenshot you can see the section of "Label encoding" in data preprocessing phase.

```
# categorical columns
categorical_columns = df.select_dtypes(include=['object']).columns

# LabelEncoder
label_encoder = LabelEncoder()
for column in categorical_columns:
    df[column] = label_encoder.fit_transform(df[column])

```

3) STANDARD SCALING:

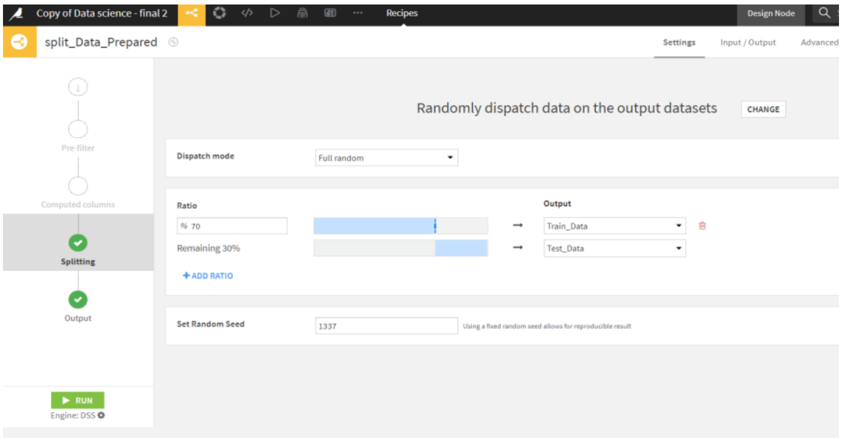
Another preprocessing method that we have implement is “StandardScaler” which is useful for classification models. StandardScaler standardizes the features of a dataset by transforming them to have a mean of 0 and a standard deviation of 1. In the following picture you can see Standard Scaling part of our code.

```
# Standard Scaling
standard_scaler = StandardScaler()
X_standardized = standard_scaler.fit_transform(X)

# dataset after data preparation
df_standardized = pd.DataFrame(X_standardized, columns=X.columns)
df_standardized['y'] = df['y']
df_standardized
```

4) SPLITTING THE DATA:

Based on our task we have splitted the dataset into 70%- for training and 30% for testing.

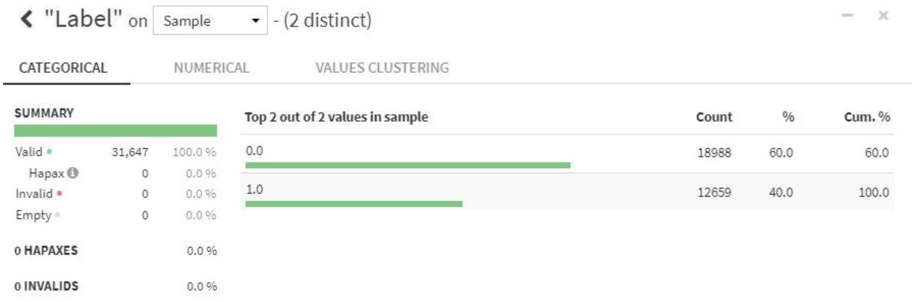


5. LOAD BALANCE ON TRAIN DATA:

For balancing the data, we have used technique known as a combination of random oversampling and random undersampling. This is a type of balancing strategy used to address class imbalance in binary classification problems. Using python recipe, we did the load balancing. We have used 60-40 ratio for balancing the data among other ratios such as 50-50, 70-30 and 80-20.

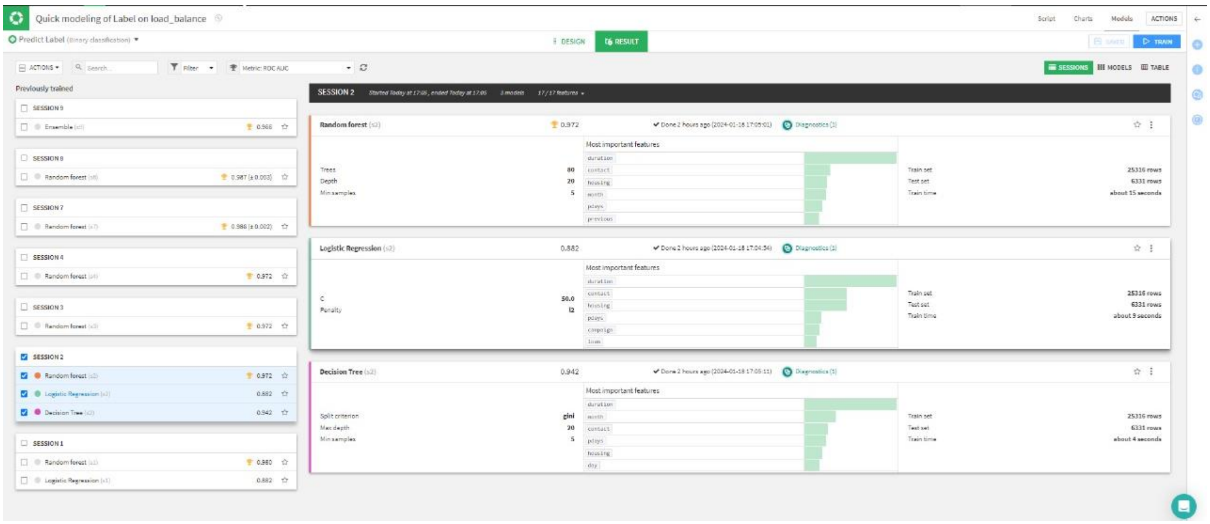
BEFORE BALANCING:

AFTER BALANCING:

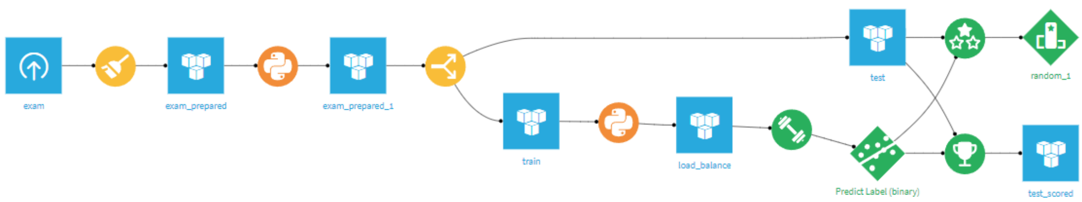


MODELLING ALGORITHMS:

After splitting the dataset now, it is the time to train a model. For Classification problems we chose Random Forest, Decision Tree and Logistic Regression. Which is shown in the Picture. The best of them was Random Forest.



FINAL FLOW :



FINAL RESULT: RANDOM FOREST ON TEST DATA

Performance metrics

Accuracy	Precision	Recall	F1 Score	Cost Matrix Gain	Log Loss	ROC AUC	Average Precision	Calibration Loss	Lift
0.879	0.486	0.764	0.594	0.060	0.290	0.922	0.599	0.121	2.417

SCORE RECIPE ON TEST DATA : (PREDECTION)

"Label" on Sample - (2 distinct)

CATEGORICAL NUMERICAL VALUES CLUSTERING

SUMMARY

Valid	13,563	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %

0 HAPAXES 0.0 %

0 INVALIDS 0.0 %

Top 2 out of 2 values in sample	Count	%	Cum. %
0.0	11992	88.4	88.4
1.0	1571	11.6	100.0

"prediction" on Sample - (2 distinct)

CATEGORICAL NUMERICAL VALUES CLUSTERING

SUMMARY

Valid	13,563	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %

Top 2 out of 2 values in sample	Count	%	Cum. %
0.0	11096	81.8	81.8
1.0	2467	18.2	100.0

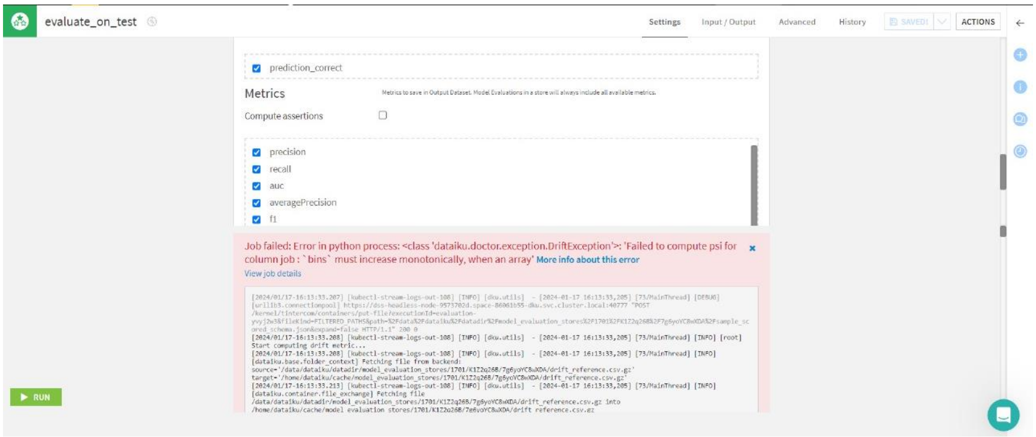
Out of 88.4% of Label data, model predicted 81.8% data for 0 value.

11.6% of Label data, model predicted 18.2% data for 0 value.

WHAT DID NOT WORK WITH OUR PROJECT:

SMOTE technique: we decided to implement SMOTE technique in Dataiku, but we could not Install imbalanced-learn library in Dataiku. It is balanced approached for undersampling and oversampling. To oversample the minority class and undersample the majority class, for example, you may use the Synthetic Minority Over-sampling Technique (SMOTE).

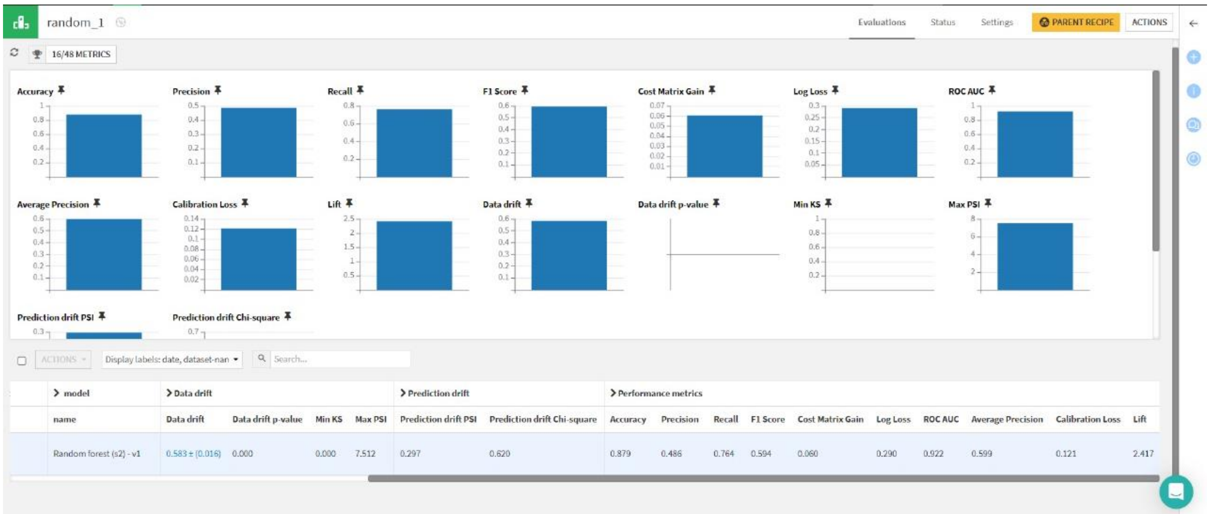
During model evaluation, we faced this error number of times. We tried to resolve it, but we were not able to tackle this problem.



HYPERPARAMETERS OPTIMIZATION:

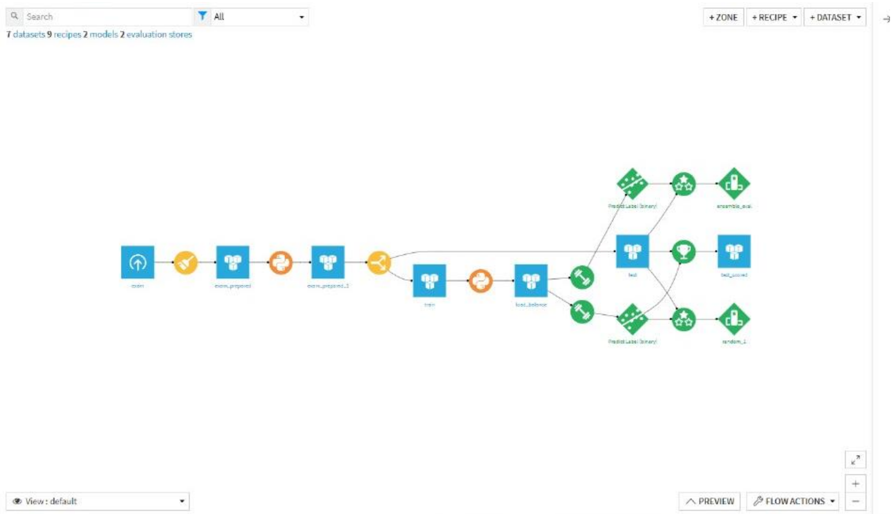
It's time to select the optimal hyperparameters following all the data preparation. The selection of hyperparameters is a crucial step in the data modelling process, since it greatly influences the machine learning model's performance. The model may underfit or overfit the training set if the hyperparameters are specified wrong, which would result in poor performance on fresh data.

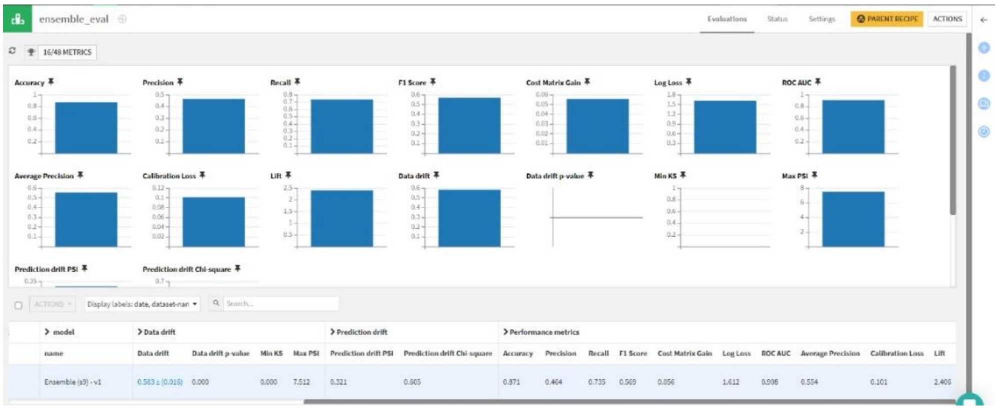
The Grid Search approach is being used to optimize parameters. We also used Random search approach, but we got the better result with Grid Search approach. We used weighting strategy as class weights. Moreover, we worked with number of trees, depth and minimum sample leaf-80-20-5 respectively, we got the best result.



ENSEMBLE METHOD:

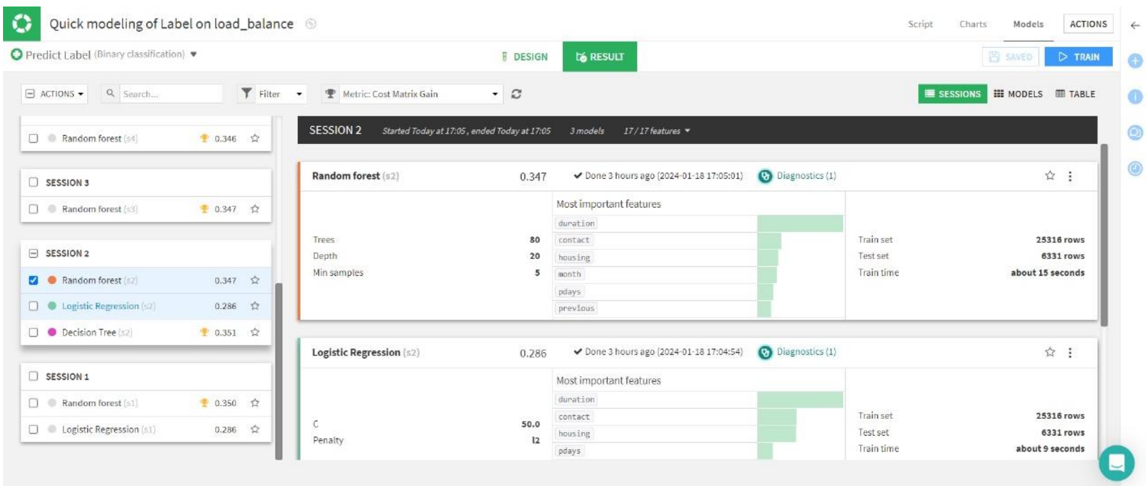
We tried different Ensemble methods, from which we got best result for Logistic staking. As per the graph, we got the worse result than the Random Forest algorithm which we chose.





TAKE IMBALANCY OF COST INTO ACCOUNT: MISSING A CLIENT IS WORSE:

The model seems to perform well when predicting the positive class (1.0) when the true value is also 1.0, as it results in a positive gain of 2,314.00. However, the model incurs a loss when predicting the positive class (1.0) but the true value is 0.0, resulting in a negative gain of −117.30.



Cost matrix

If model predicts 1.0	and value is 1.0	the gain is	1	×	2314	=	2,314.00
	but value is 0.0	the gain is	-0.3	×	391	=	-117.30
Model predicts 0.0	and value is 0.0	the gain is	0	×	3480	=	0.00
	but value is 1.0	the gain is	0	×	146	=	0.00
	Average gain per record		**0.35**	×	6331	=	2,196.70

CONCLUSION:

For this given classification problem, we got the bestter result with random forest algorithm depending on recall and ROC AUC curve evaluation metrics.