Workshop on Machine learning – Day1

ML development for classification Problems

Date: 10th July 2021

Day 1 Agenda

- **Introductions**
- **General steps to build a model**
- **Classification Models Use Cases**
- **Model Development**
  - Practical Demo using R

Dayananda Ubrangala

Senior Data Scientist
at VMWare

Experience

MARKELYTICS
Your Global Research Partner

SOL ANALYTICS
Delivering value

manthan

TATA
TATA CONSULTANCY SERVICES

Market Probe
www.marketprobe.com

Capgemini

vmware®

Publications

SmartEDA

DriveML

JOSS
The Journal of Open Source Software

General steps to build a model

Stages in Building Machine Learning Classification Model

Data Extraction & Preparation → Exploratory Data Analysis (EDA) → Feature Engineering → Building ML Models → Model Validation → Impact

Multiple Data Sources like Customer Transactions, Firmographics, etc..

Understanding which are the **important drivers**

Based on the understanding of EDA and in discussions with the product experts, **new features are engineered** which can help in predicting better model accuracy

Building Machine learning models using all the meaningful features available and using techniques to handle big data and sparsity to finally generate **scored list**

Using the scored list from the model to do a real-world validation using either Test / Control or approach agreed with the business to **call-out model success**

Performance Measurement

Data Cleaning and Feature Engineering

A Missing variable imputation

X1	X2	X3
12	A	1
NA	B	0
22	NA	NA
34	A	1

X1	X2	X3
12	A	1
23	B	0
22	A	1
34	A	1

B Removing duplicates

X1	X1	X2
1	1	0
2	2	1
1	1	0
2	2	1

X1	X2
1	0
2	1
1	0
2	1

C Cleaning column names

X1	X2	X?3
1	A	0
2	B	1
1	A	0
2	A	1

X1	X2	X3
1	A	0
2	B	1
1	A	0
2	A	1

D one hot encoding for categorical features

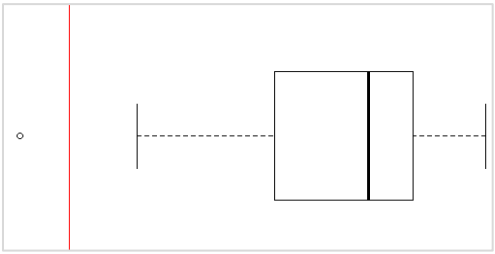
X1	X2	X3
12	A	1
23	B	1
22	C	0
34	A	1

X2_A	X2_B	X2_C
1	0	0
0	1	0
0	0	1
1	0	0

E Outlier flag and imputation

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3

X1_cap	X1_imp
0	68
0	55
1	61.6
0	62



F Bulk interactions for numerical features

X1_X3_multiple = X1 * X3

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3

X1_X3_multiple
136
220
6
186

G Frequent Transformer

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3

X2_Freq	X2_Prop
2	0.50
1	0.25
1	0.25
2	0.50

H Date Variable Transformer

X1	X2	X3
1/11/2019	A	2
12/1/2020	B	4
23/6/2018	C	6
18/9/2017	A	3

X1_month	X1_year	X1_quarter
11	2019	4
1	2020	1
6	2018	2
9	2017	3

Evaluate the model for classification problems

Confusion matrix

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

Precision tells us how many of the correctly predicted cases actually turned out to be positive

$$Recall = \frac{TP}{TP + FN}$$

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

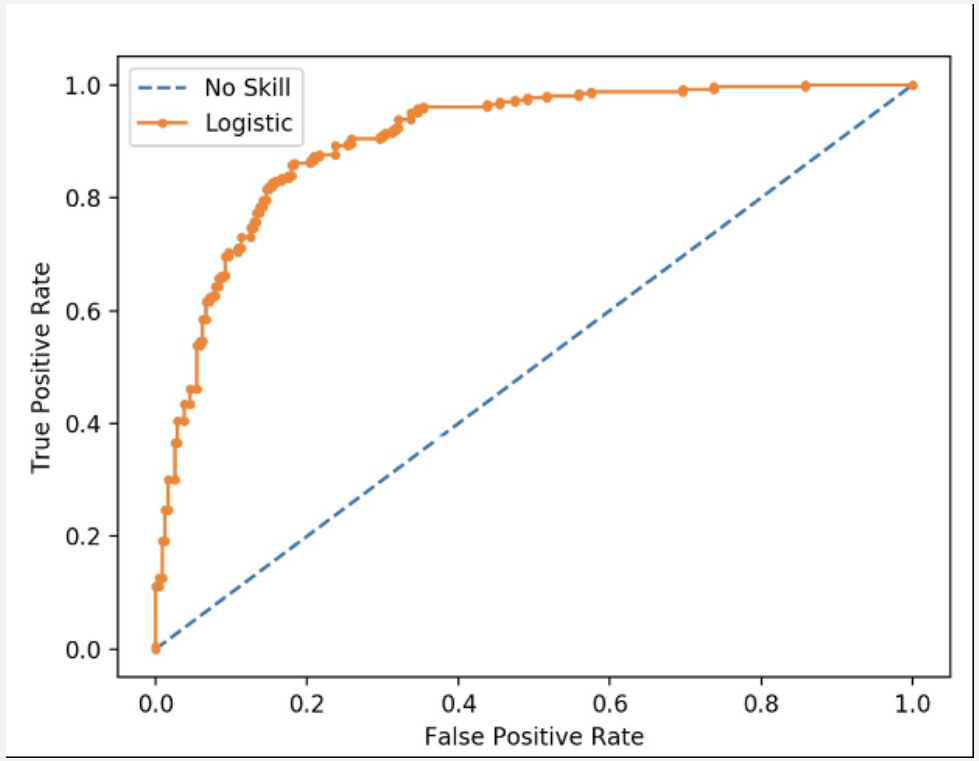
$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

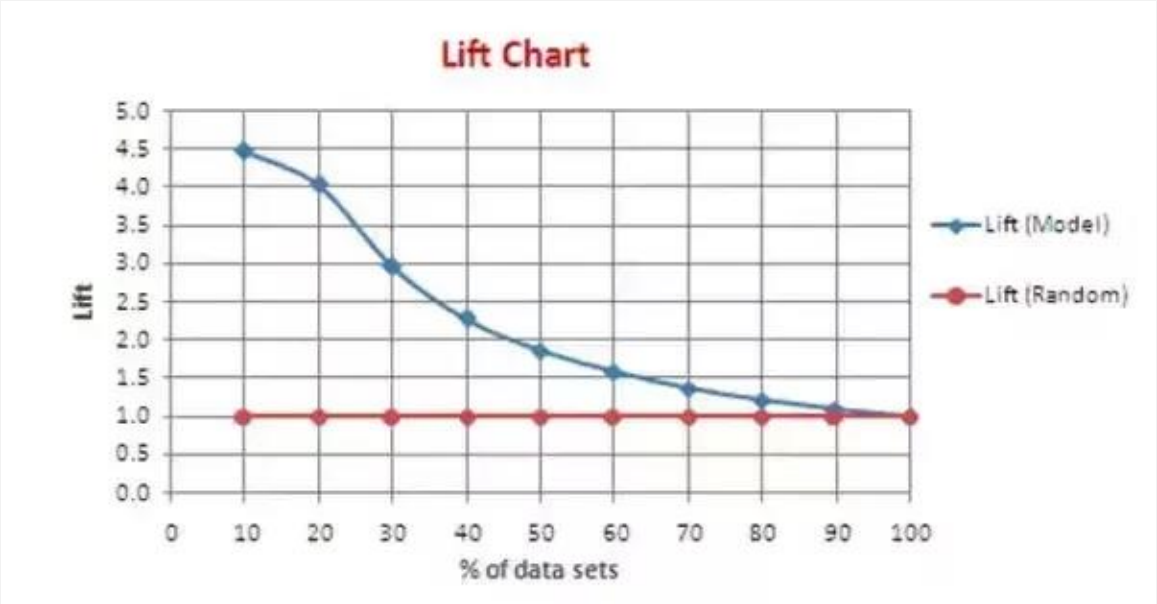
Evaluate the model for classification problems

AUC - ROC



ROC is a probability curve and AUC represents the degree or measure of separability

Lift



Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.

Classification Models Use Cases

HR Analytics: Job Change of Data Scientists

Objective

Predict the probability of a candidate looking for a new job

Data Source

Open-source Competition Data set

Evaluation

Area under the curve score

Details

This dataset designed to understand the factors that lead a person will work for the company(leaving current job) ,and the goal of this task is building model(s) that uses the current credentials, demographics, experience to predict the probability of a candidate looking for a new job or will work for the company.

Data Note

- The dataset is imbalanced so it might affect your result if you don't handle it
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality so encoding methods and techniques will help to boost models performance
- Missing imputation strategy might affect the results so it can be a part of your pipeline as well.

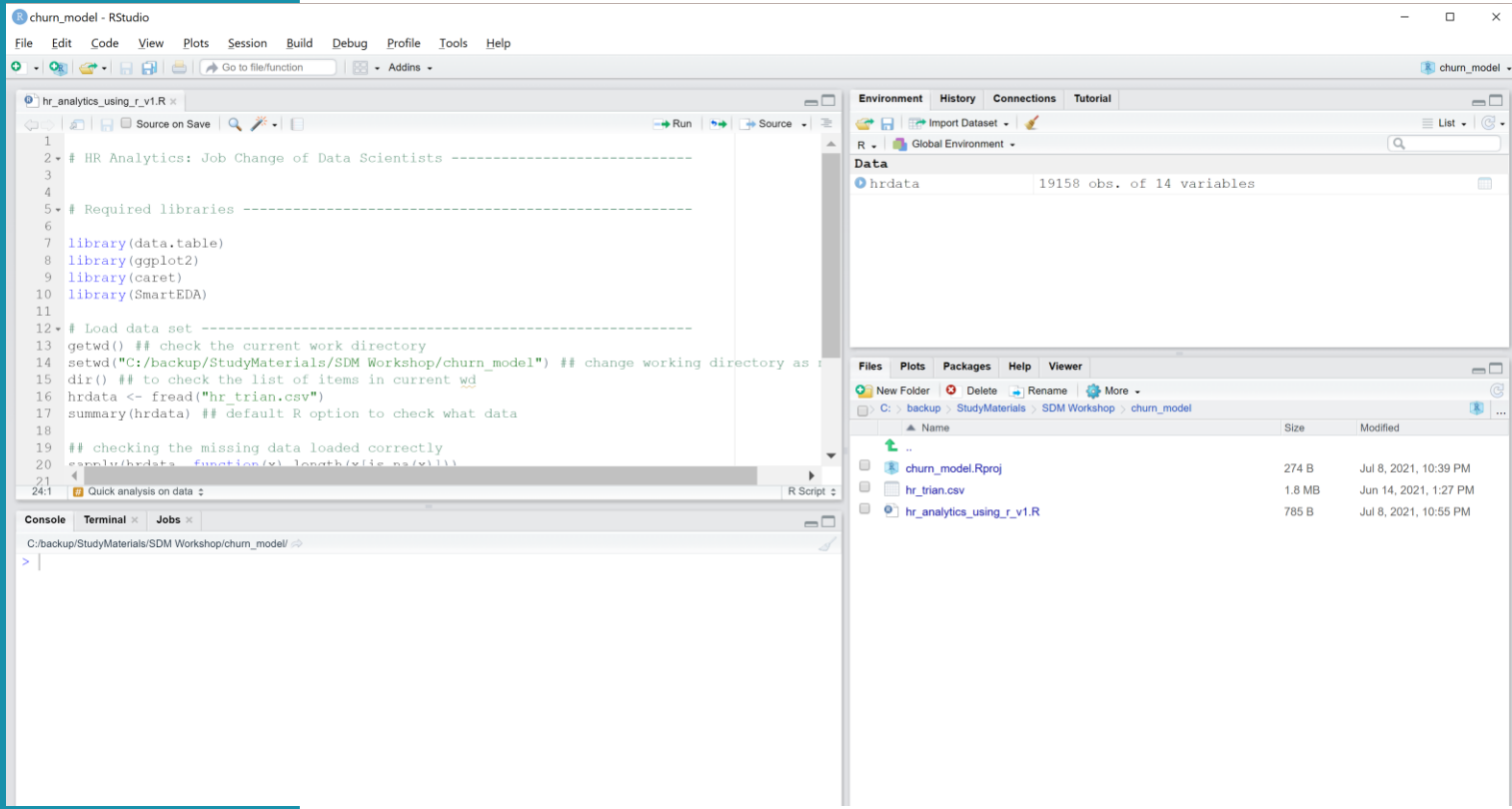
Model Development

Practical Demo using R

Logistic Regression

Random Forest

XGBoost



Q&A