International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Using Ensemble Learning Methods for Analyzing Term Deposits Subscription

Kodamagulla Kausthub[1]

Student, Department of Computer Science and Engineering, Vidya Jyothi Institute of Technology, Telangana, India

---

Abstract - *The Term deposits are generally considered to be one of the best investment strategy for those who would like to have a safe and risk-less return on their investment. The advantages of having term deposits is that it is generally simple in nature and are guaranteed by the government. Subsequently, any bank would love to market their term deposits to their customers through which they can increase their customer base and also overall revenue. It is very important for a bank to understand their clients and generally prioritize them accordingly. In order to make the banks acquire new customers and also maintain the old customer base, we propose an optimized method where we basically use ensemble machine learning methods to identify the right and clients who would love to subscribe to their bank's term deposit.*

Key Words: **XGBoost, EDA, Sci-kit-Learn, Term-Deposit, RMSE**

1.INTRODUCTION

Term deposit is usually considered to be a fixed time investment. Marketing this kind of investment is generally considered to be a tedious task for the banks as they involve understanding various tasks like client prioritization and also reaching out to them through various social media applications. In order to make the process hassle free and less chaotic, we propose a ensemble learning method which uses auto-ml technology in order to identify the right customer and accordingly understand the various performance metrics done in order to comprehend the model findings and results.

2. BACKGROUND

2.1 BANK DATA-SET CLASSIFICATION

Before performing the data exploration process, we need to understand the data-set in order to perform analysis. The data-set contains 11000 rows and 17 columns in it. The data-set is generally compiled together with the help of the bank telemarketing data-set present in kaggle repository. The classification of the data-set generally helps us to understand the various attributes like the customer's existing balance, his previous loans taken etc.

2.2 PACKAGES

In order to understand and cater to respective clients, we need to build a proper machine learning methodology which involves various steps such as data exploration, feature engineering, model validation etc. We mainly import three packages such as pandas, numpy and matplotlib. Pandas is basically used to perform data manipulation functions whereas we import matplotlib in order to visualize our findings in an intuitive manner.

2.3 EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis(EDA) is a very essential component when it comes to machine learning. Using EDA techniques, we can derive various insights and also understand our data-set. EDA helps us to pick out the imbalances, anomalies or any outliers present in our data-set.

Head(): Head function is generally used to return the first n rows mentioned in our code line. The default value for n is considered to be 5

Describe(): Describe function is basically a descriptive statistical function which helps us to summarize the data-set tendency values with the exception of Nan values.

Outliers: Seaborn plot is basically used to find out any if present outliers are properly detected.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11157	33	blue-collar	single	primary	no	1	yes	no	cellular	20	apr	257	1	-1	0	unknown	no
11158	39	services	married	secondary	no	733	no	no	unknown	16	jun	83	4	-1	0	unknown	no
11159	32	technician	single	secondary	no	29	no	no	cellular	19	aug	156	2	-1	0	unknown	no
11160	43	technician	married	secondary	no	0	no	yes	cellular	8	may	9	2	172	5	failure	no
11161	34	technician	married	secondary	no	0	no	no	cellular	9	jul	628	1	-1	0	unknown	no

11162 rows × 17 columns

Chart -1: Data-set

2.4 DATA-SET DESCRIPTION

Data-set description is an essential step while trying to understand our data. We basically perform data-set description using either .info() or .describe() function.Various sub- functions like count(), std(), mean() etc. can be properly calculated and also be stated correctly.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration
type	int	enum	enum	enum	enum	int	enum	enum	enum	int	enum	int
mins	18.0					-6847.0				1.0		2.0
mean	41.2319476796274					1528.5385235620836				15.658036194230426		371.993818
maxs	95.0					81204.0				31.0		3881.0
sigma	11.91336919221552					3225.413325946149				8.420739541006462		347.12838
zeros	0					774				0		0
missing	0	0	0	0	0	0	0	0	0	0	0	0

Chart - 2: Data-set Description

3. PROPOSED METHODOLOGY

3.1 FEATURE ENGINEERING

The Feature engineering is considered to be one of the most important steps while building an predictive model. Selecting the right features is always essential because it helps us to better model accuracy built with good parameters. In this particular data-set we basically perform various functions such as

Train_test_split: We basically split our data-set into training data and also testing data. In this particular data-set, we use 8800 rows for training and 2300 for testing purpose.

Outliers: Outliers in this data-set are essentially detected with the help of percentiles concept. We basically calculate the z-score also which can be used as a method of standardization.

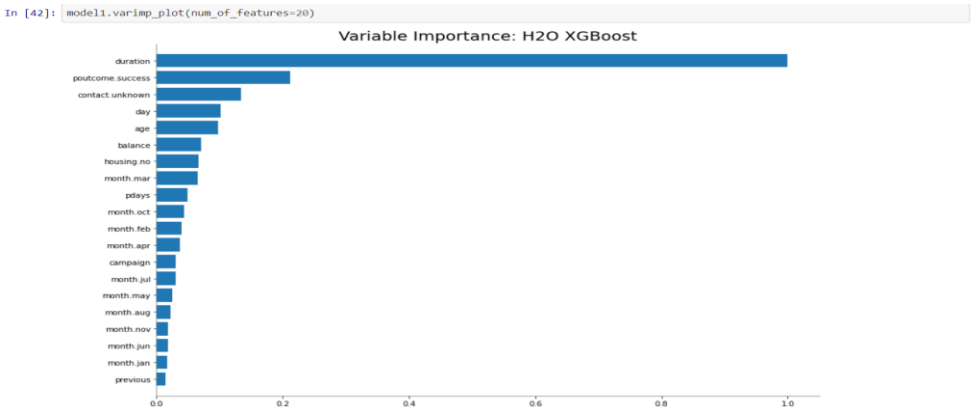

Chart - 3 - Feature Extraction

3.2 MODELLING ALGORITHM

Modelling algorithms is generally considered to the penultimate step for building an ML model. We essentially select an algorithm based upon the data-set features which we have extracted so that a relationship between the features can be established in order to understand the output. In this case, we use an ensemble machine learning method namely boosting methodology in order to build a strong classifier.

3.3 ALGORITHM EXPLANATION

3.3.1 XGBOOST

XGBoost is the algorithm which we will be using in order to predict the customer who would likely subscribe to term deposits. It is basically defined as the better boosted implementation of decision trees applied specifically for structured or tabular data. XGBoost is generally considered to be very quick in execution and also noticeable increase in model performance can be observed.The reason behind using XGBoost is

1) Better Execution Speed: XGBoost is considered to be 10 times faster than a normal classifier algorithm. It is quick in execution and also optimization of hardware is done properly

2) Better Accuracy: XGBoost generally provides better accuracy than other existing models. Even AUTO-ML uses gradient boosters which in turn uses automatic hyper-parameter tuning.

3.3.2 XGBOOST STEPS

In case of ensemble learning methods like XGBoost , we need to make sure that there is a little chance of over-fitting. In order to perform boosting methodologies, we need to make three simple steps

1. Let's take an initial model A0 which is used to predict a given target variable Y

2. A new model namely M1`is used to fit in the residuals from the previous step

3. Now A0 and M1 are combined to give A1

It can be done for n iterations until the residuals have been minimized(optimized)

$$An(x) = An\text{-}1(x) + Mn(x)$$ [Formulaic Representation]

3.4 PERFORMANCE EVALUATION

Performance evaluation is considered to be one of key steps after rendering any ML model. It helps us to understand various metrics attached to our model. Various evaluation metrics such as error rate, accuracy, log loss, rmse etc. can be found out using the evaluation metric method

```
In [40]: model1 = h2o.get_model('XGBoost_grid__1_AutoML_20200618_084430_model_12')

In [41]: model1.model_performance(test)

ModelMetricsBinomial: xgboost
** Reported on test data. **

MSE: 0.10614137575894413
RMSE: 0.3257934556723694
LogLoss: 0.34187573293235723
Mean Per-Class Error: 0.1398812389963442
AUC: 0.9234164179731832
AUCPR: 0.889088234377074
Gini: 0.8468328359463664

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.3300592824816704:
```

		no	yes	Error	Rate
0	no	2303.0	620.0	0.2121	(620.0/2923.0)
1	yes	174.0	2398.0	0.0677	(174.0/2572.0)
2	Total	2477.0	3018.0	0.1445	(794.0/5495.0)

Chart - 4 - Performance Metrics

As observed from Chart - 4 , we can see that various error rates such as mse, rmse and auc are mentioned

AUC: AUC generally refers to area under curve. It gives us the overall summarization of performance under all classification values. More the AUC, more you can distinguish between positive and negative classes. As the value of AUC is 0.92, it denotes that it is an excellent classifier.

RMSE: It measure how properly or how well the regression line fits with respect to data points.It is generally used for measuring quality of predictions of our data-set taken. As rmse value is 0.32, it shows us that model is an excellent predictor with the data-set taken for consideration.

4. CONCLUSION

As we get the various performance metrics related to our data-set, we can infer that the model has been fit accordingly and also predicts the right client who would like to subscribe to the term deposit offered by a respective bank . The model accuracy is nearly 87% observed as given above from the given confusion matrix. XGBoost also helps us to reduce the error or loss rates when compared to other bagging models. The model can be extended to better accuracy by fine tuning the hyper-parameters. While implementing it for future purposes, random forests or svm can also be used for implementing with the same data-set.

5. REFERENCES

[1] J. Guo and H. Hou, "Statistical Decision Research of Long-Term Deposit Subscription in Banks Based on Decision Tree," *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, 2019, pp. 614-617, doi: 10.1109/ICITBS.2019.00153.

[2] Huang, J., Chai, J. & Cho, S. Deep learning in finance and banking: A literature review and classification. *Front. Bus. Res. China* 14, 13 (2020). https://doi.org/10.1186/s11782-020-00082-6

[3] Bentéjac, Candice & Csörgő, Anna & Martínez-Muñoz, Gonzalo. (2019). A Comparative Analysis of XGBoost.