Customer Churn Analysis And Prediction Using Automl

N.Dinath Reddy, M.Pavan charan, Sai Hemanath,
S.Srinivas Rao, Mrs.Jayanthi.J
Computer science engineering
Kalasalingam academy of research and education
Virudh nagar,Tamilnadu

Abstract— In this paper we will try to find to analyze the data and spot if there is a pattern on why do customers leave the company by using approach called automl h20 framework it is process of applying machine learning concepts to real world problem using automation, because, our goal here is not to lose the customers. Here the telco customer data is being used to serve this paper.by using automl we are implementing several models and picking the top ones and analyzing with different evaluation methods like precision, recall. Etc.

I. INTRODUCTION

Today we are all living in a world called as competitive world in there everybody wants to be in the number one position . Especially, the big companies from different industries facing many problems to attract customers and w1th the help of data science we can solve this kind of problems .1n data science and ml is one of the facets of artificial intelligence (al) that facilitates the creation of computer systems that have the potential to learn from experience. Ml is a true requirement for the present scenario to eliminate human effort and have better automation with less mistakes. The aut0ml system has been developed to provide knowledge extraction from data. The h20 FRAME work is used for developing different aumoml models and compared the performance with the manual models. manual models have been developed by using different sampling, feature engineering and hyper parameter tuning. performance is compared using score and auc and some performance matrices ease of use.

II. LITERATURE REVIEW

Today, Customer churn is a concept that is quickly gaining importance and significance, especially in the world's leading and competing organizations. Customer churn can be defined as the group customers or subscribers that discontinue a service over a specific period. It is one of the of the central factors that determines the steady-state of customer over a fiscal year is characteristically very small In this experiment we are handling this problem using automated machine learning approach and to implement this we are using the frame work called h20.ai.

By using this, we can automate some machine learning workflows like feature selection, choosing the best hyper parameters and model selection .AT last framework will be going to generate multiple models (leaderboard) and we have to pick the top ones based on user requirements.

A. Auto ML

In this experiment we are finding whether the customer will churn in future are not.so that companies will take some steps to the retain their customers. AUTOML, automated machine learning can be used for automated machine learning work flow, which includes tuning model, selecting best hyper parameters, and feature selections.

a

b

B. Traditional / Auto ML

IN Traditional machine learning, we have to execute manual every step like data acquisition, data preparation, feature the customer pool that a business can support. One of the defining issues in building customer churn prediction models is that in practice we find the percentage of customer churn that occurs in proportion to the entire pool of engineering model section and hyper parameter tunning.etc. • But in auto ml , it provides an action to automated some steps like feature engineering ,hyper parameter tunning..etc.

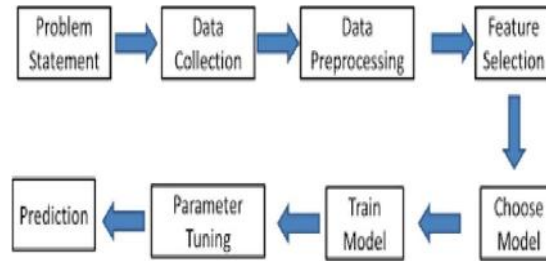Traditional Machine Learning Workflow

AutoML Workflow
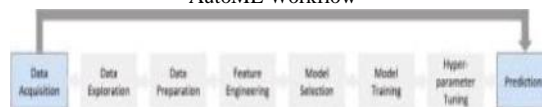
C. Methodology

For implementation this experiment we are using google we 3.7 environment is used. google cola is one of the platform for data science which has many ml libraries in python and r.

N.Dinath Reddy, M.Pavan charan, Sai Hemanath, S.Srinivas Rao, Mrs.Jayanthi.J

D. Dataset description

In this work ,we are predicting whether the customer will churn in future or no based on past data using automl .The data set is collected from Kaggle and dataset consists of following cola', 'gender', 'Senior Citizen', 'Partner', 'Dependents', 'tenure', 'Phone Service', Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tec Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method', 'Monthly Charges', 'Total Charges', 'Churn']

E. Data Preprocessing:

Outlier Detection: Box plots are used for the identification of spread, skewness and data quartiles and also outliers. Box plot indicates the information about the data by using five important parameters such as minimum, maximum, lower quantile, upper quartile, median. The data dispersiveness can be indicated from the box length.
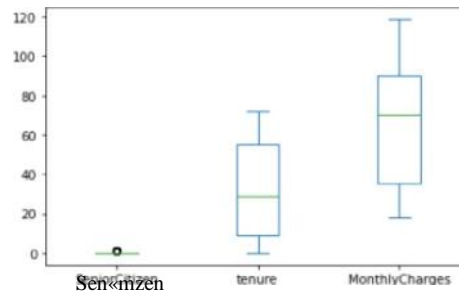

F. Label Encoding

IN this step, we have to encode some columns like Churn " , "PhoneService " , " OnlineSecurity" , "TechS upport" , " Streami ngTV","StreamingMovies","PaperlessBilling label encoding is one of encoding technique for handling categorical data. During this process each label is assigned some unique values so that model will able to find patterns.

G.   Data visualisation:

Data visualization is visual representation of data in form of graphs and plots. Especially, when we are having huge of amount data it is easy to understand and interpret the data.

N.Dinath Reddy, M.Pavan charan, Sai Hemanath, S.Srinivas Rao, Mrs.Jayanthi.J
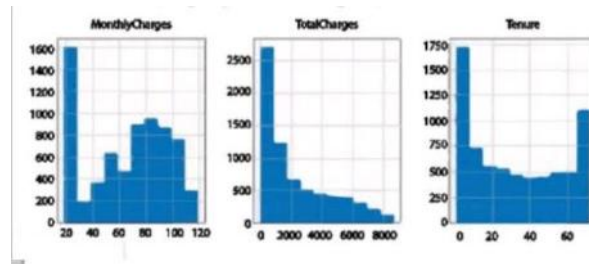
H. Models and Outcome

Implementation h20 (automl frame work) Atom: The field of ML and Al is advancing to address a variety of problems, and various approaches used to build robust models. The most important aspect of ML is the pipeline design [14], if it can be automated can tremendously help the tinkerer to use it. H20 model Each model in the H20 setting operates on clusters. After the cluster has been developed, the dataset will be loaded to start automl. The complete dataset has been separated into training and testing, used ratio is 80:20. The data set formulated is provide to the automl function. Autofluffs a whole leader board for all versions and ranks for the best running models

H20 is an automated ML model developed by H20.ai. It allows customers to acquire information from raw data without the need for experience in implementation or tuning the model. The H20 has various features and capabilities, such as clusters and flow, which helps the user to comprehend without the detail understanding. H20 is a java virtual machine capable of performing parallel computations for ML on clustering Clusters are software with one or multiple nodes. Memory is contained in a compact column format, which allows parallel reading of data. The solution is more flexible and successful because it provided better CPU efficiency and scale up flexibility. These clusters are making H20 fast. Flow is an interactive interface for code execution, text writing and plotting of the graph. H20 offers a lot of special and out-of-box features to perform quicker and more effective simulation
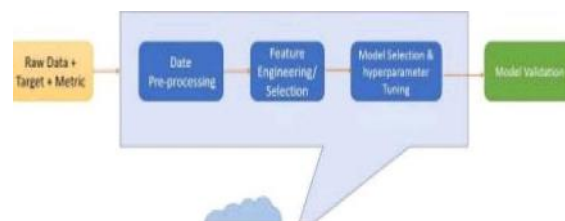
I. H20.ai automated leaderboard models

Leader board in h20 .ai auto ml provides the top performance models having high accuracy. Here, after applying automl to the data set glum (generalized linear models) model is having best auk (area under curve) .so for predicting the weather customer will churned in future are not we are considering glm_l model in this case

N.Dinath Reddy, M.Pavan charan, Sai Hemanath, S.Srinivas Rao, Mrs.Jayanthi.J

predict	No	Yes
Yes	C.130665	0.869335
Yes	C.486019	0.513981
No	C.961202	0.0387979
No	C.826175	0.173825
Yes	C.480061	0.519939
No	0.82035	0.17965
Yes	C.159316	0.840684
No	0.9772	0.0228002
No	0.87115	0.12885
No	C.947038	0.0529623

Confusion Matrix (Act/Pred) for max f1 @

		No	Yes	Error	Rate
0	No	593.0	191.0	0.2436	(191.0/784.0)
1	Yes	49.0	207.0	0.1914	(49.0/256.0)
2	Total	642.0	398.0	0.2308	(240.0/1040.0)

```
1b.head()
```

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
GBM_1_AutoML_2_20220315_123431	0.849914	0.40964	0.654929	0.233083	0.365693	0.133658
GLM_1_AutoML_2_20220315_123431	0.842793	0.417418	0.665437	0.247068	0.367472	0.135036
GBM_2_AutoML_2_20220315_123431	0.840273	0.42228	0.649105	0.233576	0.372281	0.138593
XGBoost_3_AutoML_2_20220315_123431	0.837458	0.424138	0.648611	0.247792	0.372373	0.138661
XRT_1_AutoML_2_20220315_123431	0.836458	0.423974	0.654493	0.24386	0.370377	0.137179
GBM_3_AutoML_2_20220315_123431	0.835714	0.42922	0.629456	0.242271	0.375185	0.140764
GBM_4_AutoML_2_20220315_123431	0.833333	0.43503	0.640644	0.24742	0.376724	0.141921
XGBoost_2_AutoML_2_20220315_123431	0.830825	0.443965	0.61107	0.232619	0.381236	0.145341
XGBoost_1_AutoML_2_20220315_123431	0.830348	0.440719	0.620101	0.248409	0.379697	0.14417
DRF_1_AutoML_2_20220315_123431	0.620415	0.468494	0.633844	0.25991	0.378249	0.143073

J. Results

In classification kind of problems, roc and auk are the best performances measurement methods. When we need to check the model performance we can use auk and roc curves. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting correctly
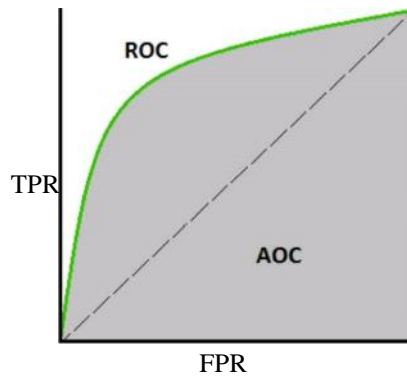
Predicted matrix and confusion matrix:

Conclusion and future work

Atom has been developed for saving time and energy for routine tasks in ML pipelines. The top performing models are provided by the leader board with respect to its increasing values of accuracy and AUC (area under curve) values as well as with the decreasing error parameters. Atom tools produce fair performance outcomes across a wide variety of datasets, but in multi-class classification scenario, it suffers from poor performance.

N.Dinath Reddy, M.Pavan charan, Sai Hemanath, S.Srinivas Rao, Mrs.Jayanthi.J

REFERENCES

[1]   Pradeep B, S. V. Rao, S. M. Puranik and A. Hegde, "Analysis of Customer Churn prediction in Logistic Industry using Machine Learning" , Int. J. of Scientific and Res. Publications, vol. 7, no. 11, 2017. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, PP.6V3.

[2] R. Mattison, The Telco Churn Management Handbook, 2001 .K. Elissa, "Title of paper if known," unpublished.

[3] Forhad N, Hussain MS, Rahman RM, "Churn analysis: Predicting churners", In Digital Information Management (ICDIM), 2014 Ninth International Conference, 2014.

[4] Dahiya K., Bhatia S., "Customer churn analysis in telecom industry", Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference, pp. 16, 2015. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] N. Lu, H. Lin, J. Lu, G. Zhang, "A customer churn prediction model in telecom industry using boosting", IEEE Trans. Ind. Informat., vol. 10, no. 2, pp. 1659-1665, 2014.

N.Dinath Reddy, M.Pavan charan, Sai Hemanath, S.Srinivas Rao, Mrs.Jayanthi.J