INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1470)
Available online at: https://www.ijariit.com

Importance of Feature Selection in Model Accuracy

Pranjal Rawat
rawatpranjal111@gmail.com
Uttaranchal University, Dehradun, Uttarakhand

Nitin
santynitin7@gmail.com
Uttaranchal University, Dehradun, Uttarakhand

Sameer Dev Sharma
samsharma11@gmail.com
Uttaranchal University, Dehradun, Uttarakhand

ABSTRACT

As a dimensionality reduction strategy, feature selection attempts to select small set of most important features from primary features by eliminating obsolete, data-redundant and non-relevant noisy features. This process of choosing a set of the original variables such that a model based on data containing only these features has the simplest output is known as feature selection. Feature Selection eliminates over-fitting, increases model efficiency by removing redundant functions, and has the added benefit of maintaining the primary feature representation, resulting in improved accuracy. Good learning efficiency, results into higher machine learning model accuracy, lower cost of computation, and efficient model accuracy, is typically the product of feature selection. Recently, researchers in the area of computer vision, deep Learning, data mining, and other fields have shown that several feature selection algorithms resulted in the efficiency in their work through computational theory and research. This paper aims to examine the importance of feature selection in model accuracy. Feature selection is critical for various reasons, which include simplicity, performance, computational efficiency, and accuracy. It is often used in both supervised and unsupervised learning scenarios. These strategies can help boosting the productivity of various machine learning algorithms, as well as coaching. Feature selection decreases learning time and increases data consistency and comprehension.

Keywords: *Feature selection, Over-fitting, Supervised learning, Unsupervised learning, Computer vision*

1. INTRODUCTION

The size of this data has been growing exponentially over a few years. According to a research conducted by UK based mobile network company GiffGaff. The data usage has grown from 9.4 billion GB in 2016 to over 67 billion GB by 2021. This is approx 720 percent increase in only a Five-year interval. On account of the rise in the quantity of raw information, the dimensions of sample information and its features also raise that are utilized in many Machines Learning software like Computer Vision, Text mining, etc... Because of the existence of a quite high number of non-relevant dimensionality data, they could make our machine learning model slow and inaccurate. Therefore, it's essential and important to choose relevant features from a huge range of dimensional data to raise the productivity of the machine learning algorithm. Feature Selection is capable of picking attributes on your information (for instance, columns in tabular data) that are relevant to this predictive modeling problem you're working on.

Feature selection is also termed as variable and characteristic selection. Feature selection differs from dimensionality reduction. Both approaches try to decrease the number of characteristics from the data-set, but a dimensionality reduction system does this by producing various new combinations of features, in compared to this feature selection methods tends to include and exclude characteristics within the data without altering them.

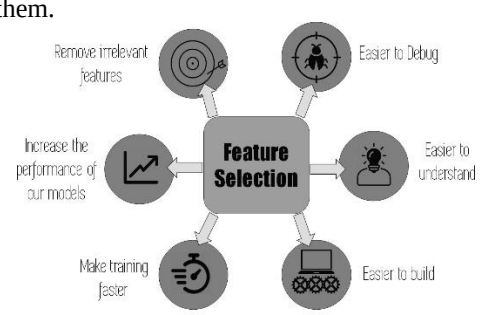

Fig. 1: Feature Selection

The feature selection procedure is categorized into three major types, based on the availability of the label data. Supervised methods, Semi-supervised methods, and Unsupervised methods. In the Supervised selection technique, all the data are labeled. When a number of our data is labeled, we could use a Semi- supervised selection technique which may use both labeled and unlabeled data for optimum outcomes. Unsupervised feature selection can be used where the data isn't labeled.

2. METHODOLOGIES USED IN FEATURE SELECTION

Based on the searching strategies, feature selection is categorized into three major methods:

2.1 Filter Method

These methods are usually used as preprocessing steps. Features are chosen according to their score in several statistical evaluations for their correlation with the resultant variable. Therefore, the option of statistical measures is extremely dependent on the variable data types. Frequent datatype contains

2.2 Wrapper Method

Wrapper methods, try to use a set of features and then train a machine learning model using these sub-features. Dependent on the knowledge we draw in the former version, we opt to include or remove attributes from the subset. These methods are often computationally extremely costly.

- Numerical Variables
- Categorical Variables

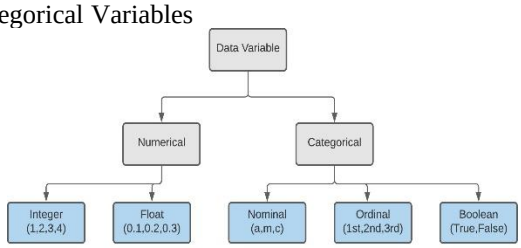

Fig. 2: Data Variables Types

Filter feature selection methods are usually calculated by only one input variable with the target variable. Filter feature selection methods are also known as uni-variate statistical measures. This implies that any fundamental interaction between various input factors is not evaluated in the filter feature selection methods.

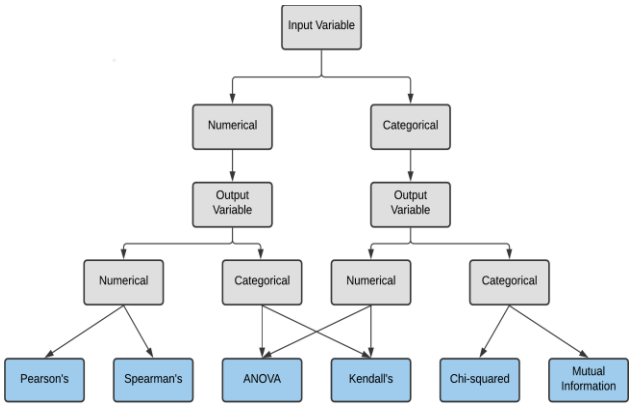

Fig. 3: Feature Selection Technique

There is numerous statistical theory test that is utilized to compute a correlation between input and the target variable. The tests have been performed depending on the kind of input (numerical, categorical) and the output factor (numerical, categorical).

The filter methods don't eliminate multicollinearity. Thus, you should deal with the multicollinearity of attributes too before training of machine learning models for the data.

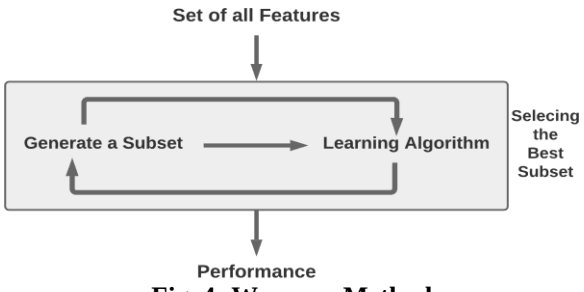

Fig. 4: Wrapper Method

2.3 Embedded Method

Embedded method contains characteristics of both filter as well as wrapper methods. Embedded method do feature selection while training of the machine learning model.

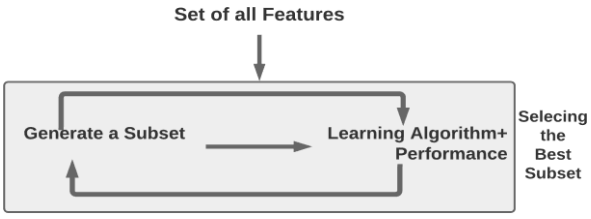

Fig. 5: Embedded Method

3. MACHINE LEARNING MODEL ACCURACY

Machine learning model precision is the measurement used to ascertain which version is better at identifying patterns and relationships between factors in a dataset based on the input signal, or training, information. The greater a version can generalize to examine hidden information, the greater forecasts and insights it could create, which then deliver greater business value.

Business firms use machine learning models to produce sensible business decisions, and also much more precise model results lead to better choices. The price of mistakes can be enormous, but maximizing model precision mitigates that price. For instance - A false positive cancer identification, prices the hospital and the individual.

The DataRobot automatic machine learning platform utilizes top open-ended calculations to empower its users to develop exceptionally precise, highly interpretable versions. It thoroughly analyzes the accuracy of its versions with Factors like target leakage that might inhibit model precision and therefore Negatively affecting the decision-making procedure.

4. BENEFITS OF PERFORMING FEATURE SELECTION BEFORE DATA MODELING

- Improve Accuracy: Less dimensionality of non-useful, misleading data lead to increase in the accuracy of our machine learning model.
- Reduce Training Time: Fewer data attributes reduces algorithms complexity resulting in less algorithm time to run resulting in faster model training.
- Reduce Overfitting: Less redundant data resulting in less similarities between train and output variable, thus reduces overfitting of data models on test datasets.
- Remove Collinearity: Those variables which share common characteristics are termed as collinear variables. These variables may decrease the model's availability to learn and can also decrease the interpretability of the machine learning models, and thus decrease the performance of model on the test datasets.
- Remove Missing Values: A relatively simple choice of feature selection is removing missing values. Missing values make our model under-fitted and decrease accuracy.

5. EXPERIMENTS

We concluded an extensive experiment on the supervised learning dataset. For this on Feature Selection, research we used "***Titanic – Machine Learning from Disaster***" publicly available dataset from ***Kaggle***. The competition is simple: use machine learning to create a model that predicts which passengers survived the 1912, Titanic shipwreck. Let's have a look at the sample data.

The Train data consists of 12 columns and 891 rows. The rows are as follows: ***PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked***. To have a look at the dataset go to our Kaggle notebook - Titanic-Feature_Selection with Exploratory Data Analysis for more detailed, Exploratory Data Analysis of the data-set.

Sample Data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Fig. 6: Train_Sample_data

The objective of the competition is to create a Machine Learning model in order to predict the Survival rate of the people's who survived in the 14 Apr 1912 incident of Titanic ship. The model is to be created on the *Train data-set* and tested on the *Test data-set* which contains all the same Features as train dataset except ***Survived*** column.

We have applied various data analysis techniques to our dataset such as – Exploratory Data Analysis (EDA), Feature Selection, Feature Engineering, Checking for outliers and missing values. With the use of various visualization tools such as Heatmap, histogram, bar-graph, line-chart, co-relation matrices and many other tools we can easily visualize the relation between the features and can select the best appropriate features for our model.

In our study it is found that various features are correlated to each other. This relationship can be used to create new features with feature transformation and feature interaction. We also used target encoding because of high correlations with survived features.

For the dataset we used ***Random Forest classifier*** model for best optimum results. By applying the test dataset on the model, we found that the model achieve accuracy of almost 80% , which is considered to be a good score, but we can still improve our classifier with a clever grouping approach and ensembling models.

6. CONCLUSION

This paper helps to gain much information regarding feature selection, methods used for feature selection, the importance of feature selection in machine learning models, and the benefit of feature selection. From the paper, it is clear that feature selection is one of the first important steps before training our model using data. Feature selection can play a vital role inaccuracy of the machine learning model.

ACKNOWLEDGMENT

I would like to thanks all the authors for their time and support in conducting this research. I am also thankful to Kaggle for providing such a great environment for the data scientist to works with various publicly available datasets.

REFERENCES

[1] Jianyu Miao, Lingfeng Niu, A Survey on Feature Selection, ITQM, 2016.

[2] L. Wolf, A. Shashua, Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight- based approach, The Journal of Machine Learning Research 6 (2005) 1855–1887.

[3] P. S. Bradley, O. L. Mangasarian, Feature selection via concave minimization and support vector machines., in: ICML, Vol. 98, 1998, pp. 82–90.

[4] J. G. Dy, C. E. Brodley, Feature selection for unsupervised learning, The Journal of Machine Learning Research 5 (2004) 845–889.

[5] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th international conference on Machine learning, ACM, 2007, pp. 1151–1157.

[6] S. Doraisami, S. Golzari, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization and Similarity, 2008.

[7] I. Guyon, A. Elisseeff, An introduction to variable and feature selection J. Mach. Learn. Res., 3 (2003), pp. 1157-1182.

[8] G. John, R. Kohavi, K. Pfleger, Irrelevant Features and the Subset Selection Problem, International Conference on Machine Learning (1994), pp. 121-129.

[9] Esra Mahsereci Karabulut, Selma Ayse Ozel, Turgay ibrikci, A comparative study on the effect of feature selection on classification accuracy, 2012, Pages 323-327.

[10] M. Ramaswami and R. Bhaskaran, A Study on Feature Selection Techniques in Educational Data Mining, 2009.

[11] Marginalia, Global mobile data usage to increase by 720% by 2021, 2017.

[12] Jason Brownlee, How to Choose a Feature Selection Method for Machine Learning, 2019.

[13] Jason Brownlee, How to Calculate Feature Importance with Python, 2020.

[14] Robinson Spencer, Fadi Thabtah, Neda Abdelhamid, exploring feature selection and classification methods for predicting heart disease, 2020.

[15] Pranjal Rawat, Titanic - Feature_Selection with Exploratory Data Analysis, 2021.

