Feature Importance Scores

Mr. Akshay Prasad Arun Kumar UOG Mr. Mohan Sangli & Mr. Anish Ravishankar

ABSTRACT

A dataset is a combination of features and target. Not all features in a given dataset is important. It is necessary to understand the contribution of each and every feature in the dataset towards the decision making of the machine learning model. The proposed experiment looks out for the best possible way to treat the datatypes and get the best importance scores for each and every feature in the dataset. It also focuses on the difference between the methods to treat the features and the results obtained by them thus formulating a procedure to treat similar datatypes or features in the future. Also an attempt was made to understand the effect of overfitting on the feature importance scores.

I. Introduction

Feature importance scores are an important aspect of any regression models. These scores tell the importance of any given feature in the dataset.

The data is always a combination of numerical and Categorical datatypes. Thus, direct evaluation of importance scores of such data is not advisable. Such type of data should be preprocessed before feeding it into any machine learning model and getting their importance scores.

There are various techniques that deal with tuning of the data which are Label Encoding, One Hot Encoding, Standardization, Normalization and to name a few.

The Categorical data is treated with Label Encoder and One Hot Encoder while the numerical data is treated with mathematical transforms such as Normalization and Standardization.

II. Need of Experiment

Treating of different types of data with different techniques may or may not offer same feature importance scores. It is very important to understand the variation of feature importance scores of different techniques.

Correlation in the dataset is also a huge factor that changes the importance scores significantly for the correlated columns. An attempt to understand whether a particular model (in the regression category) is able to distinguish between correlated columns is important.

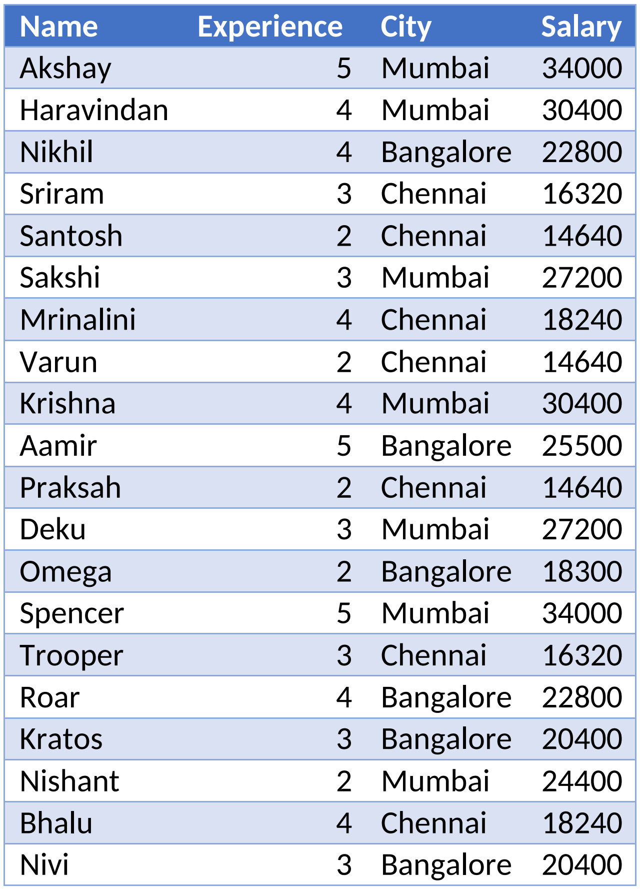
Six regression models were taken into account and the readings were recorded.

1. Random Forest Regressor
2. XG Boost Regressor
3. Support Vector Regressor
4. Decision Tree Regressor
5. Ridge Regressor
6. Lasso Regressor

III. Insights on the Dataset

Salary dataset was used to differentiate between feature importance scores of Label Encoder and One Hot Encoder.

Every city in the dataset has a base salary fixed. The incentives that is given to the employees are uniform throughout all cities.



Name	Experience	City	Salary
Akshay	5	Mumbai	34000
Haravindan	4	Mumbai	30400
Nikhil	4	Bangalore	22800
Sriram	3	Chennai	16320
Santosh	2	Chennai	14640
Sakshi	3	Mumbai	27200
Mrinalini	4	Chennai	18240
Varun	2	Chennai	14640
Krishna	4	Mumbai	30400
Aamir	5	Bangalore	25500
Praksah	2	Chennai	14640
Deku	3	Mumbai	27200
Omega	2	Bangalore	18300
Spencer	5	Mumbai	34000
Trooper	3	Chennai	16320
Roar	4	Bangalore	22800
Kratos	3	Bangalore	20400
Nishant	2	Mumbai	24400
Bhalu	4	Chennai	18240
Nivi	3	Bangalore	20400

Fig. Salary Dataset

The dataset used to study correlation is the automotive dataset
(Link : https://drive.google.com/open?id=15JutvEUhvMmzBRPcGqAZVRmEiJBWSKPB).

An extra column 'KW' is added to the dataset which is a derivative of the 'horse power' column in the dataset. This column has direct correlation with the 'horse power' column in the dataset.

IV. Observations

To distinguish between One hot encoder and Label Encoder, the importance scores of the former are the arithmetic sum of the later. More precise the dataset, less difference between the importance scores of the two techniques.

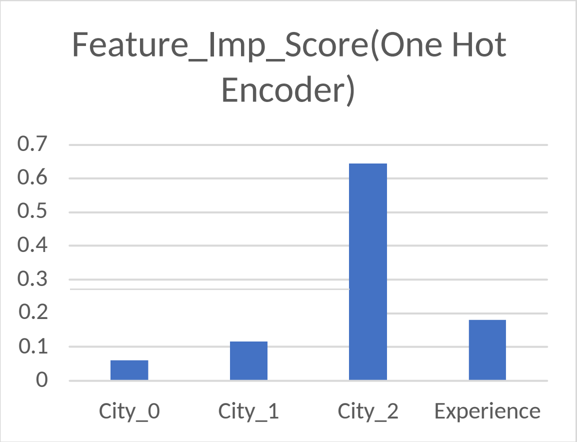


Fig. Importance Scores (OHE)

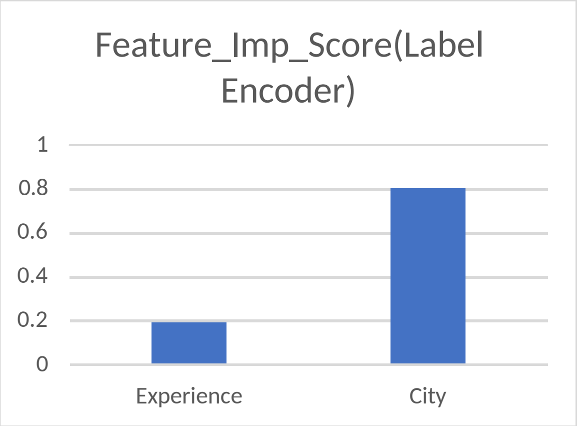
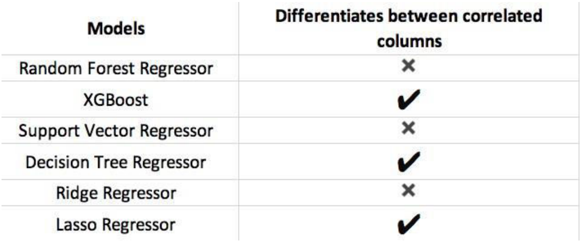


Fig. Importance Scores (LE)

It is evident from the two observations that the sum of feature importance scores of the one hot encoder features in equal to the feature importance score of the label encoded feature.
Label encoder can be directly used until Euclidean distance is not calculated. One hot encoder is required when there more than one columns in Euclidean distance based models.

The capability of the six regressor models to distinguish correlation in the dataset are as follows:



Models	Differentiates between correlated columns
Random Forest Regressor	✗
XGBoost	✓
Support Vector Regressor	✗
Decision Tree Regressor	✓
Ridge Regressor	✗
Lasso Regressor	✓

Fig. Distinguishing capability of models

Overfitting any model in some cases changes the feature importance scores but the changes in the feature importance scores are not so significant. Overfitting just memorizes the given dataset but doesn't change the decision-making algorithm significantly.

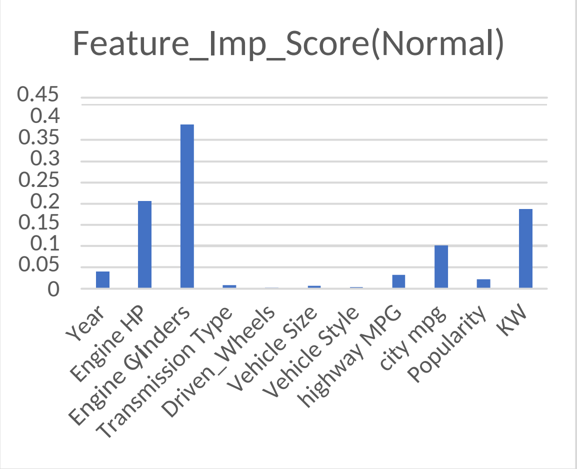


Fig. Importance Scores (Normal)

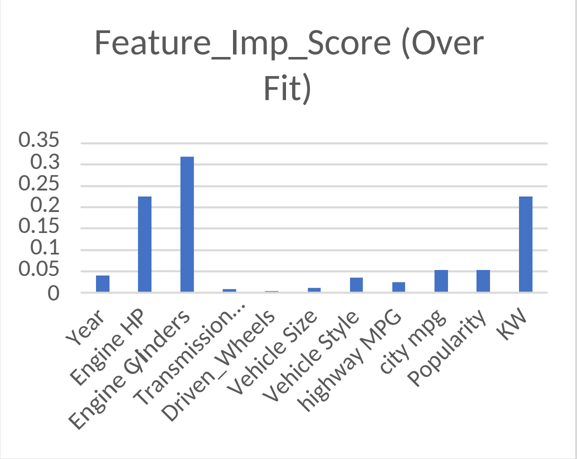


Fig. Importance Scores (Over fit Model)

V. Conclusion

The findings of the above experiments reveal that Label Encoder and One Hot Encoder give the same importance scores provided the data is precise. While dealing with more than one categorical columns with one hot encoder, the dummy variable trap should be taken care of.

Correlation in the dataset is unwanted and it increases variance in the data and hence, the efficiency of the model reduces which is not ideal. In such cases the correlated columns should be deleted and then the data should be fed into the model.

Overfitting the model will not change the importance scores significantly. The slight change in the feature importance scores will not change the algorithm of decision making of the model for that particular problem.