Sales Prediction

Vikash Kumar

27th July 2019

Contents

Introduction

Problem Statement

We have got a problem to predict the sales of the items with given different details of items such as where it sold from, size, type and eight more variables. If a company want to predict the sales according to some of the criteria, before investing more into it. It should know which item is on sales or high demand and can make that item abundant at that place at that time.

Data

The data I received having missing values and needs some pre-processing before implementing the algorithms to it. Some encoding must be done and than after encoding, removal or imputation of missing values are done. And finally feature selection is done.

The characteristics of data is not favorable we have to apply data pre processing. After the pre processing of data becomes favorable and handy to implement the algorithms on the top of it.

Methodology

Pre-Processing

Data is having missing values and some outliers, to deal with that we first try to encode the data in the form of factors which will influence the target variable. Factor data is well structured data so to predict the final outcome or the sales prediction we need the independent variables to be the form of factors as the prediction becomes accurate and informative.

Data encoding

Data encoding is done to encode different factors of a different variable to form in a single dimension. In our data set we try to encode most of the variable and make in the useful form and then try to put a model on the top of it.

Missing value analysis

The data I received having missing values in the two variables, which could I either deleted or imputed with values. So the data is limited or they are precious in accuracy of our model. So we try to impute with values using knn or statistical techniques.

Normalization

Data should be normalized for the proper modelling and to enhance the accuracy of the model. The goal of **normalization** is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require **normalization**

Modelling

I used different regression model to check, which model performs well and then selecting the model on the basis of their score and evaluation metrics. I have used two evaluation metrics to determine the accuracy of model i.e. R^2 square and RMSE value. To measure the performance of the regressions three standard regression metrics are used: Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Both metrics are calculated for both regressor types. For comparison RMSE is used and R^2 for parameter tuning. “The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.”

We uses different modelling techniques to check the accuracy of the model on the given sample data. On the basis of evaluation metrics we check the model predictions and its accuracy level.

Model evaluation

We evaluate model on the basis of R2 score and RMSE value and then select the model as required. We got different evaluation score for each evaluating model and then we select the model which is having high accuracy and less errors in prediction.

Results obtained from Linear Regression:-

```
Training Score :  0.6952720126422746
Validation Score :  0.694894141104267
Cross Validation Score :  0.6934739186933732
R2_Score :  0.5709561243806104
MAPE output: 6.396575
RMSE : 0.562445
```

Results obtained from Support Vector Regression:

```
R^2_Score SVR: 0.725618
RMSE SVR: 0.533375
MAPE output: 5.936216
```

Conclusion

As expected the tuning of the parameters can improves the performance. Parameter tuning with grid search can improve the performance even further after we apply different regression model on it.

The tuned SVR beats all the model by far and gives 72% coefficient.

A coefficient of determination of more than 72% is a decent result for the SVR regressor. All the other models performance is not satisfactory and cannot be selected in the predictions of rental bike count. We have to stick to the tuned SVR model for that as if now.

In coming future we can apply some more regression algorithm and try to calculate the accuracy and its coefficient.

Reflection

We have used different regressor model to check which model is performing well and which is having coefficient value is higher. Higher the coefficient value better the model. We have created decision tree model, linear regression model of two types and a support vector regression. We can also implement other models such as DNN regressor which is also a good model for high amount of data, which can also be used here to check if it is working fine when tuned properly.

Improvement

The coefficient of determination of the regressors could be increased by additional iterations in training and the number of folds in the cross validation, at the expense of computing time. Of course, there are also other regressors available that might perform better on this particular dataset. For example, a wide and deep learning algorithm might be a better-performing alternative. We can also different hyperparameter tuning techniques such as randomized search and many other techniques to tune a model.

References

> http://www.capitalbikeshare.com

> http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

> http://freemeteo.de/wetter/

> http://dchr.dc.gov/page/holiday-schedules

> http://scikit-learn.org/stable/

> https://www.tensorflow.org

> https://www.tensorflow.org/versions/r0.9/tutorials/linear/overview.html#large-scalelinearmodels-with-tensorflow

>https://www.tensorflow.org/versions/r0.9/api_docs/python/contrib.learn.html#DNNRegressor

> http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

> http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR

> http://scikit-learn.org/stable/modules/sgd.html#regression

> http://scikit-learn.org/stable/modules/svm.html

> http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

> http://scikit-learn.org/stable/modules/sgd.html#regression

> http://www.vernier.com/til/1014/

> https://github.com/tensorflow/skflow/blob/master/g3doc/api_docs/python/

estimators.md

> http://scikit-learn.org/stable/modules/generated/

sklearn.grid_search.GridSearchCV.html

>http://scikitlearn.org/stable/modules/generated/sklearn.grid_search.RandomizedSearchCV.html

>http://scikitlearn.org/stable/modules/generatedsklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error

> http://scikit-learn.org/stable/modules

Sales Prediction