Pipeline

Introduction

The general background of ML, and the scenario.

Define the objectives

- Demensionality reduction
- Clustering
- Regression
- Classification

Download Data

EDA

Brief introduction to the data set

- How many features
- Convert the categorical feature into numerical feature
  - Label encoder
  - One-hot encoder

Check The Dataset

- Clean
  - Move On
- Not Clean
  - Missing Values
    - Discard Rows With Missing Values
    - Imputation (0 Or Mean Value)
  - Useless Columns
    - Ignore Some Columns
  - Etc..

Descriptive statistical analysis

- Correlation matrix
- Pairplot
- Boxplot
  - Outlier detection

Feature Selection

Filter Method

Filtering and taking only the subset of the relevant features. The model is built after selecting the features. The filtering is using correlation matrix, and it is most commonly using.

At first, plotting the Pearson correlation heat-map and see the correlation of independent variables with the target variable. We could only select features which has correlation of above 0.5 (taking absolute value) with the target variable.

- A value closer to 0 implies weaker correlation (exact 0 implying no correlation)
- A value closer to 1 implies stronger positive correlation
- A value closer to -1 implies stronger negative correlation

Hence we can drop all other features apart from these. However this is not the end of the process. One of the assumptions of linear regression is that the independent variables need to be uncorrelated with each other. If these variables are correlated with each other, then we need to keep only one of them and drop others. Next, checking the correlation of selected features with each other. This can be done either by visually checking it from the above correlation matrix.

Wrapper Method

A wrapper method needs one machine learning algorithm and uses its performance as evaluation criteria. It needs to feed the features to the selected Machine Learning algorithm, and based on the model performance to add/remove the features. This is an iterative and computationally expensive process but it is usually more accurate than the filter method.

Backward Elimination

At first, we feed all features into the model. Then checking the performance of the model and then iteratively remove the bad performing features one by one till the overall performance of the model comes in acceptable range.

It would use p-value to evaluate the feature performance. If the p-value is above 0.05, we can remove the feature, otherwise we can keep it.

Here we are using OLS model which stands for "Ordinary Least Squares". This model is used for performing linear regression. The P-value of features are shown below[x].

As we can see that the variable 'XXX has highest p-value of XXXXX which is greater than 0.05. Hence we can remove this feature and build the model once again. This is an iterative process and can be performed at once with the help of loop. This approach is implemented below, which would give the final set of variables which are xxx,xxx,xxx,xxx.

RFE (Recursive Feature Elimination)

The Recursive Feature Elimination (RFE) method works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance. The RFE method takes the model to be used and the number of required features as input. It then gives the ranking of all the variables, 1 being most important. It also gives its support, True being relevant feature and False being irrelevant feature.

We took LinearRegression model with all features, and RFE gave feature ranking as above. Then we need to find the optimum number of features, for which the accuracy is the highest. We do that by using loop starting with 1 feature and going up to the number of features. We then take the one for which the accuracy is highest.

As seen from the result, the optimum number of features is XXX. We feed XX as number of features to RFE and get the final set of features given by RFE method, as follows.

Embedded Method

Embedded methods are iterative in a sense that takes care of each iteration of the model training process and carefully extract those features which contribute the most to the training for a particular iteration. Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold.

We can do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes its coefficient and convert it into 0. Hence the features are removed if its coefficient = 0, and keep the rest of features.

Normalization

Advantage

- Normalization dramatically improves model accuracy.
- Normalization gives equal weights or importance to each variable so that no single variable steers model performance in one direction just because they are bigger numbers.
- For example, clustering algorithms use distance measures to determine if an observation should belong to a certain cluster. "Euclidean distance" is often used to measure those distances. If a variable has significantly higher values, it can dominate distance measures, suppressing other variables with small values.

Disadvantage

- Normalization compresses data within a certain range, reduces the variance and applies equal weights to all features. The model lose a lot of important information in the process.
- For example, normalization lease absolutely no traces of outliers. The model treat outliers as noise and it needs to get rid of them as soon as possible. But outliers are real data points, once we lose that to get a better model, we may lose some important information.

Model Selection

Classification

- Binary classification
- Confusion matrix

Regression

- MSE(mean square error)
  - RMSE is sensitive to outliers, and the variance makes the influence of outliers more obvious if there are fewer outliers in the data.
- MAE(Mean absolute error)
  - MAE is not as sensitive to outliers. It is suitable for the case where there are many outliers in the data. Because it is correlated with median.

Clustering

- Kmeans
- How to select best k in kmeans?
- The Silhouette Method
  - The silhouette value measures how similar a point is to its own cluster compared to other clusters.
  - The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.
  - The Silhouette Score reaches its global maximum at the optimal k. This should ideally appear as a peak in the Silhouette Value-versus-k plot.
  - There is a clear peak at k = xx. Hence, it is optimal.

Dimensionality Reduction

- Select X Features, Class Label
- Pca(x), Plot The Z1,z2 Data, Colour Using Labels !% Var Captured, Scree Plot
- Reduce Data To 2 Features(z1,z2).
- Compare Performance Of Classifier With The Model Without Dimension Reducing.

Apply The Models

Simple Model

- KNN
  - What Is The Best Value Of K? Plot The Performance Of Different K.
- Evaluation
  - Training/test Split 10 Fold Cross Validation
  - Confusion Matrix
- Analysis

Complex Model

- MLP
  - Num Of Hidden Units/layers
    - The number of hidden units and hidden layers are important to the performance of model.
  - Learning Rate
    - The learning rate is large at first, then I would try smaller value, if the loss is still good and without sacrificing speed of training.
  - Optimizer
    - I choose the SGD, because it is faster than batch gradient descent. And Batch gradient usually costly. Although SGD may never converage to the minimum, the parameter β would around the minimum of J(β). It is still reasonably good approximations to the true minimum.
  - Do You Need Multiple Trials?
    - Comparing the performance curves of the model with different hyperparameter settings is a very straightforward approach. But it is quite time consuming.
- Analysis

..

Discussion

Take away message

Conclusion

- Data cleaning
- Feature selection
  - Filter method is less accurate. It is great while doing EDA, it can also be used for checking muilti co-linearity in data.
  - Wrapper and Embedded methods give more accurate results but as they are computationally expensive, these method are suited when we have lesser features.
- Evaluation metrics
  - RMSE is sensitive to outliers, and the variance makes the influence of outliers more obvious if there are fewer outliers in the data.
  - F1 score is suitable for the dataset with imbalanced distribution.