Feature Engineering Project 1

Chi-Squared (Categorical Variables)

Chi-Squared filter method is a univariate method that can be used to find the correlation between categorical feature and the target categorical variable. The result is combined with statistical method such as p-value tests to estimate the dependence of Target variable on the feature.

The formulae to compute Chi-square value is,

Chi-square = SUM ( O - E )2 / E

where, O is Observed and E is the expected value of the feature occurrences for the given target.

The Chi-squared value is shown for the sample data provided,

(using level of significance as .05, if p-val < .05 then Reject Null hypothesis (H0) )

| Features | Chi-Squared | P-value | Comment |
| --- | --- | --- | --- |
| Ba | 1.45514077e+04 | 0.00000000e+00 | Reject H0, Feature has strong relevance |
| Mg | 1.00983604e+04 | 0.00000000e+00 | Reject H0, Feature has strong relevance |
| K | 3.17203921e+03 | 0.00000000e+00 | Reject H0, Feature has strong relevance |
| Al | 1.69969035e+03 | 0.00000000e+00 | Reject H0, Feature has strong relevance |
| Na | 4.31125331e+02 | 5.78184560e-91 | Reject H0, Feature has strong relevance |
| Ca | 3.21532547e+02 | 2.34313064e-67 | Reject H0, Feature has strong relevant |
| Fe | 2.16400033e+02 | 8.77084237e-45 | Reject H0, Feature has strong relevance |
| Si | 1.10195984e+01 | 5.09927872e-02 | Reject H0, Feature has strong relevance |
| RI | 9.40575393e-03 | 9.99999545e-01 | Accept H0, Feature is not relevant |

Order of features by relevance,

Ba, Mg, K, Al, Na, Ca, Fe, Si. We drop RI feature.

### Information Gain Filter Method

In this method, we measure the entropy of Target variable with respect to each feature, and calculate Information gain. And we choose features with high values.

In general entropy is given by the following formula,

If Y is the target class discrete random variable, and Xi represents feature vectors.

H(Y) = Sum(-[(P(y) log2 p(y)]) & the information gain is given as,

I(X, Y) = H(Y) - H(Y|X), where H(Y|X) is entropy of Y given X.

| Features | IG | Comment |
| --- | --- | --- |
| al | 0.43780369 | Select this feature |
| mg | 0.39867584 | Select this feature |
| ba | 0.35728425 | Select this feature |
| k | 0.3560591 | Select this feature |
| na | 0.28429635 | Select this feature |
| ca | 0.25001787 | Select this feature |
| si | 0.0899743 | Reject this feature |
| Fe | 0.04655316 | Reject this feature |
| refractive\_index | 0.02806945 | Reject this feature |

Order of features by relevance,

Al, Mg, Ba, K, Na, Ca, Si, Fe. We drop RI feature.

### K-Nearest Neighbour Wrapper Method

Wrapper method uses feature subset selection procedure to select the most relevant features. Most commonly used algorithms is Forward selection. This method takes care of relation between two feature and its impact on the Target variable.

With forward selection we get the following 6 best features - **(‘na', 'mg', 'al', 'k', 'ba', ‘fe')**

These features are selected based on accuracy score (RMSE) of the model.

The RMSE value for selecting num features us shown below.

RMSE value for k = 1 is 3.1452039190341368

RMSE value for k = 2 is 2.6457513110645907

RMSE value for k = 3 is 2.8147960713126134

RMSE value for k = 4 is 2.7259437321355757

RMSE value for k = 5 is 2.4651415936864716

**RMSE value for k = 6 is 2.436895881744138**

RMSE value for k = 7 is 2.5328609546093555

RMSE value for k = 8 is 2.3336996049521552

RMSE value for k = 9 is 2.4242365967295925

You could easily see, lowest test RMSE is achieved at k=6 and it further starts increasing. (the indexes are offset by 1).

