**Feature selection technique**

**1.Univariate Selection**

It is statistical test.it selected those features that have the strongest relationship with the output variable.

I used selectedKbest library where k value selected number of best attribute in given dataset.

Suppose k=10, so it selected best 10 attribute. Basically selectKbest extract top features.

Based on the chi squared formula I calculate some score with respect of each and every feature and target output

**Chi Squared formula**

**=**

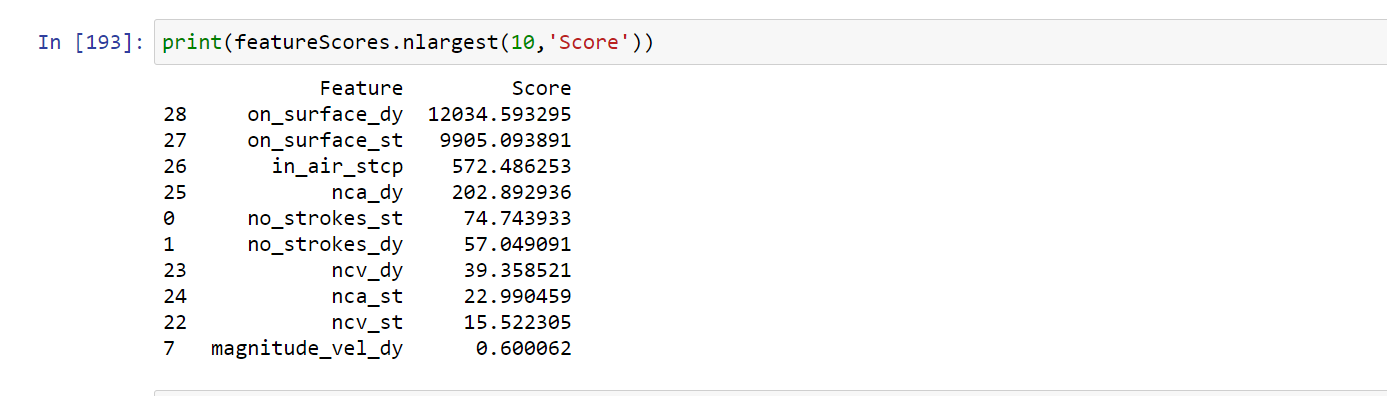
Where,

= chi squared

= Observed value

= Expected value

**Selected Feature**

****

**Accuracy**

|  |  |
| --- | --- |
| **Algorithm Name** | **Result** |
| XGBClassifier | 87.5 |
| RandomforestClassifier | 87.5 |
| LogisticRegression | 80.33 |
| ExtraTreeClassifier | 87.5 |

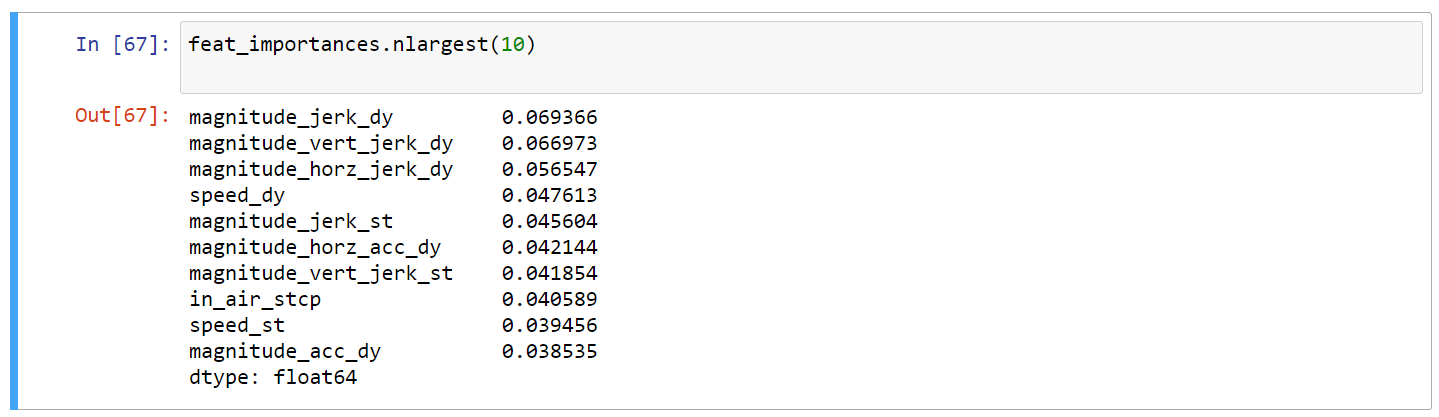
**2.Feature Importance**

By using the feature importance property of the model can get the feature importance of each feature in the dataset. It’s gives score for each feature in the data set. The higher the score more relevant output variable.

Feature importance is an inbuilt class that comes with tree based classifier

I will used ExtraTreeClassifier for extracting the top features of the dataset. It builds multiple trees and splits nodes using random subset of feature.

**Selected Feature**

****

**Accuracy**

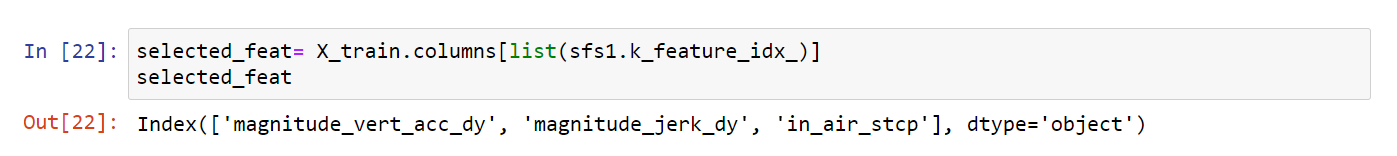
|  |  |
| --- | --- |
| **Algorithm Name** | **Result** |
| XGBClassifier | 87.5 |
| RandomforestClassifier | 87.5 |
| LogisticRegression | 80.33 |
| ExtraTreeClassifier | 93.75 |

**3. Step Forward Selection**

Starts with the evaluation of each individual feature and selects the best performing result. Best performing are selected based on Auc[[1]](#footnote-1), prediction accuracy.

I used Random forest classifier to feature selection. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

**Selected Feature**

****

**Accuracy**

|  |  |
| --- | --- |
| **Algorithm Name** | **Result** |
| RandomforestClassifier | 93.75 |

**4. Mutual Information**

Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

The function relies on nonparametric methods based on entropy estimation from k-nearest neighbor’s distances.

The mutual information between two random variables X and Y can be stated formally as follows

**I(X ; Y) = H(X) – H(X | Y)**

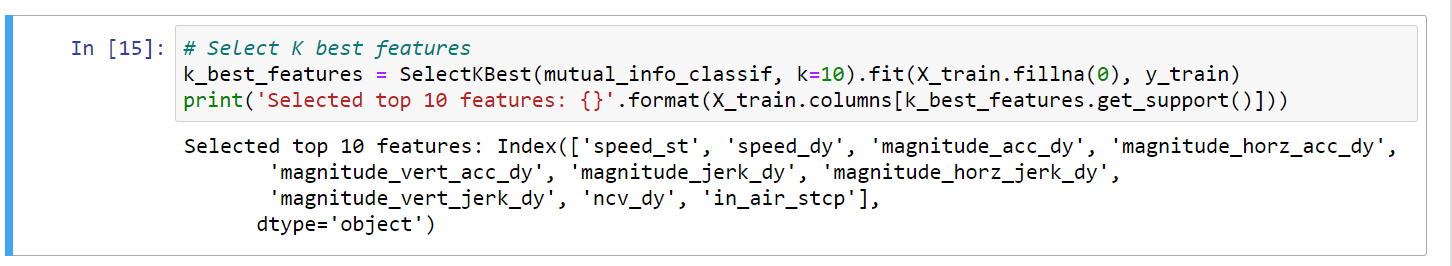
Where

I(X; Y) = Information of X with Respect to Y

H(X) = Entropy of X

H(X| Y) = Entropy of X with Respect To Y

**Selected Feature**

****

**Accuracy**

|  |  |
| --- | --- |
| **Algorithm Name** | **Result** |
| XGBClassifier | 93.75 |
| RandomforestClassifier | 87.5 |
| ExtraTreeClassifier | 87.5 |

**5. with All Feature**

|  |  |
| --- | --- |
| **Algorithm Name** | **Result** |
| XGBClassifier | 93.75 |
| RandomforestClassifier | 87.5 |
| LogisticRegression | 80.33 |
| ExtraTreeClassifier | 93.75 |
| SupportVectorClassifier | 81.25 |

**Features :**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| No of Stroke | Each letter has a specific direction,length and curvature relative to the other strokes in a letter. |
| Speed | Absolute magnitude of velocity |
| Velocity | The rate of change position with time. |
| Acceleration | The rate of change velocity with time. |
| Jerk | The rate of change of pen acceleration with time |
| Grip angel | Angel the pen is held relative to the writing surface |
| In air time | Time the pen is lifted off the writing surface during the task |

1. The Area Under the Curve (**AUC**) is the measure of the ability of a classifier to distinguish between classes [↑](#footnote-ref-1)