Alex Wolf
5/10/17

Exploratory Data Analysis Notes

Observing Data in DF

- Col Names
  - print(df.info)
- Val type
  - print(df.info)
- What entries look like
  - print(df.head()/tail())
- Size of DataFrame
  - print(df.shape)

Plots

- Scatter Plot
  - Check if linear decision boundary
  - Look at skewness

EDA Functions

- Outliers
  - Doesn't make difference regardless of Sample Size

What to Look for/add

- Regression
  - High dimensional Data
  - Feature Correlation
  - Function to remove features with poor correlation
  - Regression Equation for Outliers
- Classification
  - Normal Distribution of Data
  - Discrete Variables
  - Noise in Data
    - Test with Models
  - Dimensionality of Data

- Outliers-CHECK
- Imbalanced Data
- NaN Count
- Data Skewness

Functions

Visual EDA:

scatterMatrixPlot(isCategorical, dfX, dfY=None, diagonal='kde')
-plot scatter matrix
- if categorical maek iscategorical as true and enter DfY for scatter coloring
-diagnoal
-'kde'- density plot of vars
-'hist'- histogram of vars

correlation_matrix(dfx)
-plot correlation matrix

Statistical EDA Data Analysis:

outliers(points, stdThresh=3.5, removeOutliers=False):
-checks outliers in a data set with threshold with MAD criterion
-return Boolean array of indexes
-code from StackOverflow
-Set remove outliers to true to remove them

printClassImbalance(dfY)
-for categorical Data only

skewness(dfX, removeBadSkew=False, absGoodSkewThresh=2)
-prints skewness of each collumn
-set removeBadSkew to remove collumns with skewness above or below the absGoodSkewThresh