1/18/2019

Vedika Ahuja

HW 2 Write-up

My pipeline-library file ("ml-pipeline/read_and_clean.py") does the following, with the indicated functions:

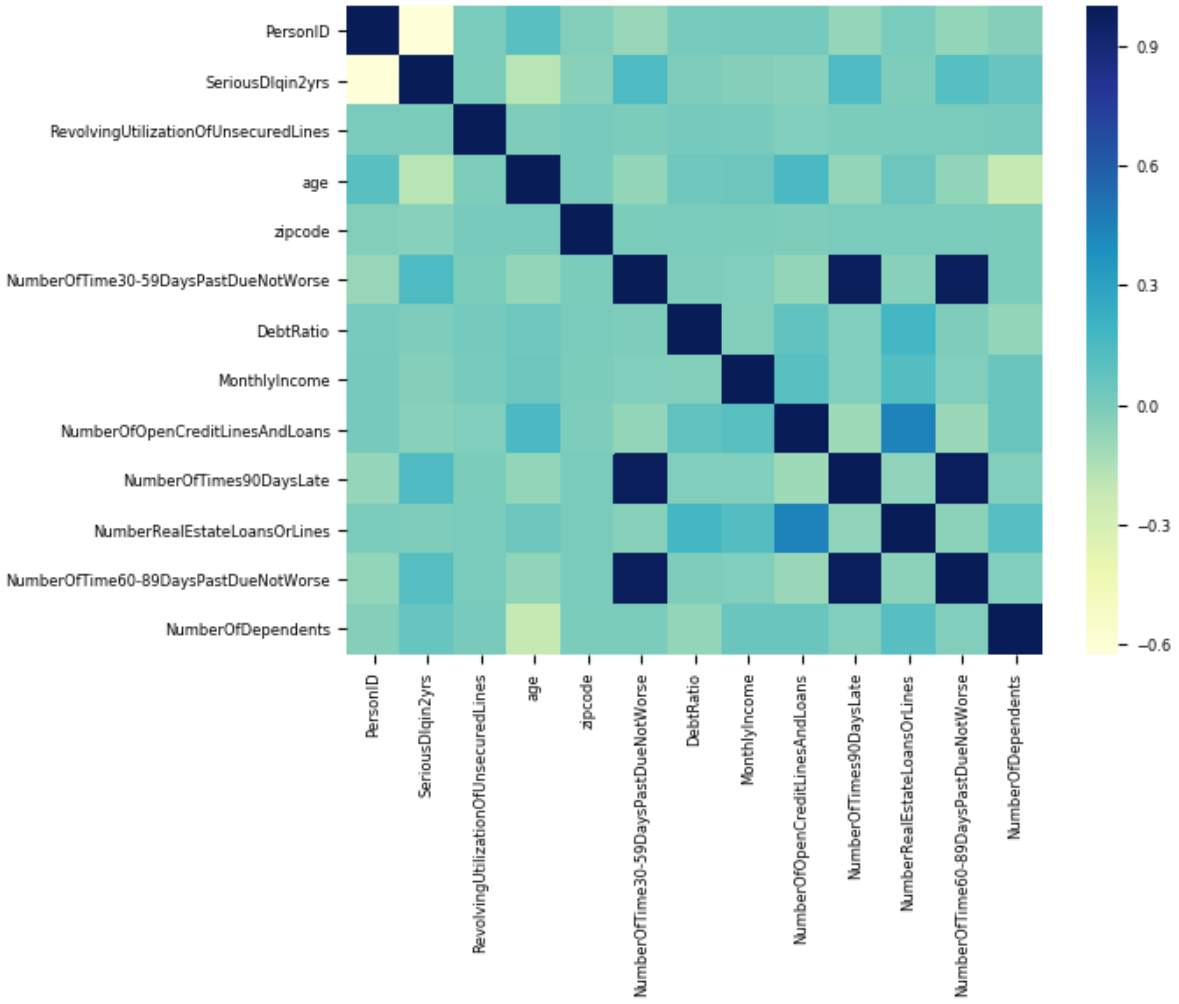
- Reads in the dataset (read_dataset)
    - This function can take a dictionary with the variable names and datatypes specified
- Conducts exploratory analysis and prints some of the results out and exports the matplotlib charts into a folder, with the following functions:
    - describe_data
    - distributions
    - correlations
- Fills missing values with the median value of the variable (fill_missing_w_median)
    - The function changes the dataframe given
- Discretizes continuous variables (discretize_cont_var)
    - This function by default cuts the data into quartiles, meaning a fourth of the data falls in each bucket. The user can specify how many quantiles they would like to split the variable into. The values of the discretized variable is the quantile number (1, 2, 3, and 4 for quartiles). The function adds the variable with the prefix "_cat" to the dataframe given
- Creates dummy variables (create_dummies)
    - The function returns the dataframe with the dummy variables joined to the inputted dataframe.
- Trains a logistic model (train_logistic_model)
    - Creates an instance of a model class, using the l1 penalty method.
    - Trains a model the data given
- Evaluate Logistic Model
    - Takes a dataframe, the prediction features, and the actual target variable. Creates an accuracy score for the logistic mode, which is the number of correct predictions over the total number of observations.
    - Predictions are 1 if the probability score is greater than .5, and 0 if it's less than .5. In a later iteration I will allow a user to choose the probability score threshold.

The script build_model_script.py can run from the command line, and when run applies all the functions in the pipeline-library to the "credit-data.csv" file. The function ultimately returns the accuracy score of the logistic regression model.

The results of the script are:

Data exploration:

- The mean of SeriousDlqin2yrs target variable is 16.1, so 16.1% of the time people experienced 90 days past due delinquency or worse
- Of the 41,016 rows, MonthlyIncome is missing 7974 and NumberOfDependents is missing 1037
- The correlation map of all the variables with each other is:



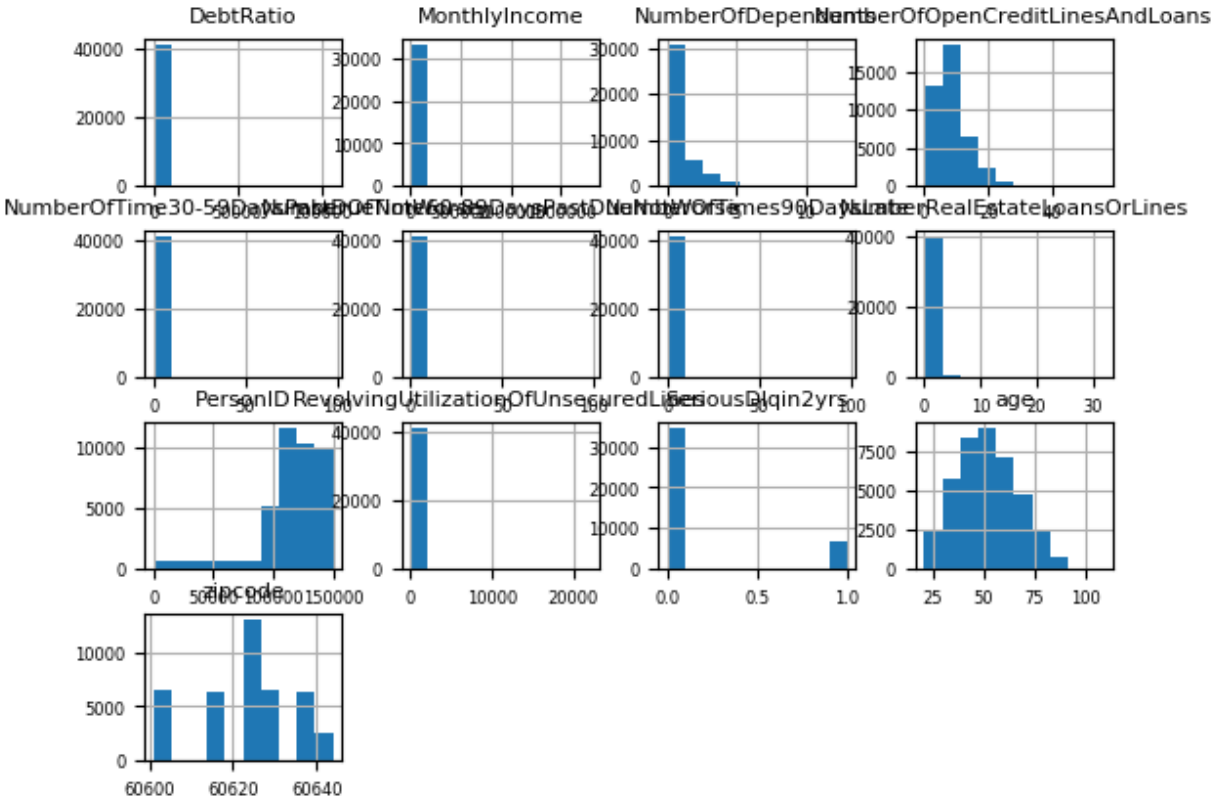
From this table we can see that the NumberOfTime30-59DaysPastDueNotWorse is highly correlated with NumberOfTimes90DaysLate and NumberOfTimes60-89DaysPastDue, along with a few other pretty intuitive correlations around being late over certain periods.

The 5 most positively and negatively correlated variables with SeriousDlquin2yrs are:

```
5 most postivively correlated variables with target variable

SeriousDlqin2yrs                        1.000000
NumberOfTime30-59DaysPastDueNotWorse    0.149334
NumberOfTimes90DaysLate                 0.139609
NumberOfTime60-89DaysPastDueNotWorse    0.121886
NumberOfDependents                      0.065708
Name: SeriousDlqin2yrs, dtype: float64

----------------------------------------
5 most negatively correlated variables with target variable

PersonID                         -0.622739
age                              -0.173728
zipcode                          -0.045051
NumberOfOpenCreditLinesAndLoans  -0.039898
MonthlyIncome                    -0.032810
```

The most meaningful (and somewhat obvious) correlations are between the 3 variables describing lateness of payment within 2 years and SeriousDlquin2yrs.

The distribution of every variable is:



And finally the script returns the score of .840, which tells us that the logistic model predicts the correct number (0 or 1) for the target variable SeriousDlquin2yrs 84% of the time on the full dataset. It is important to note that currently the model classifies each observation of 0 if the probability calculated is <.5, and 1 if the probability is >.5.