Homework 2
Shuyu Zhou

GitHub Link:
https://github.com/szhou12/MachineLearning4PublicPolicy/tree/master/pa2

1. Read Data: (refer to Read.py)

2. Explore Data: (refer to Explore.py; all mentioned output files refer to results folder)

- The summary statistics is provided in *summary_stats.csv*. The scatter plot that shows correlation between variables is shown in *scatter_matrix.png* where the diagonal is the density plot of each variable. The numeric value of correlations between variables is provided in *correlation_matrix.csv*. The outliers for each variable are provided in graphs whose names have the format as *Outlier-<variable name>.png*. The distributions of variables are provided in graphs whose names have the format as *Dist-<variable name>.png*.
- The summary statistics (Table 1) shows that about 16% of the sample experienced 90 days past due delinquency or worse. The average age of the sample is 51.7, the average monthly income is 6579, and the average debt ratio is about 331. These three indicate that the sample is centered on middle-aged group who are below middle class and have monthly debt 300 times more than their monthly gross income. The average number of dependents is less than one, so it implies that the sample is centered on persons living in small households.
- The distributions of variables are shown to be highly left-skewed whereas the distribution of age is more likely to be asymptotically normal.
- The correlation matrix (Table 2) indicates that experiencing 90 days past due delinquency or worse is mostly positively correlated with the number of times borrower has been 30-59 days past due but no worse in the last 2 years but mostly negatively correlated with age.

3. Pre-Process Data: (refer to Preprocess.py)

- Table 1 shows that 19.4% (7974 out of 41016) of data points have missing values for Monthly Income and 2.5% (1037 out of 41016) of data points have missing values for Number of Dependents. I filled in missing values by its sample mean in the sense that the sample mean is supposed to better predict the missing value.

4. Generate Features/Predictors: (refer to Preprocess.py)

- Notice the high skewness of continuous variables in the sample. Before discretizing data, I apply log transformation to variables that give high skewness score.
- I provide two methods to discretize a continuous variable:
    1) Discretize by normalization: discretize the data by how many standard deviation it's far from the mean.
    2) Discretize in quantile-base: discretize the data by ranking (0-25%, 25%-50%, 50%-75%, 75%-100%). Note due to the fact that some variables have too many 0 values, which will fail to discretize in this way. When the failure happens, the program will switch to normalization instead.

3) Note that variable *SeriousDlqin2yrs* itself is a dummy variable. Hence, there is no need to discretize.

- Based on different discretizing method is used, the creation of dummy variable changes accordingly:
    1) If discretize by normalization, then the criterion is 95% confidence interval (i.e. set dummy true if a data is within 2 standard deviation from its mean; set false otherwise).
    2) If discretize in quantile-base, then set upper 50% true and false otherwise.

5. Build Classifier: (refer to Build.py)

- Decision Tree classifier is selected in this assignment.

6. Evaluate Classifier: (refer to Evaluation.py; all mentioned output files refer to results folder)

- Split data into 70/30 where the test data take up 30% of the sample.
- Take the dummy variable of each variable once at a time as the target variable and the rest variables as attributes.
- Apply accuracy to get training set accuracy score and testing set accuracy score respectively in 1-step depth, 3-step depth, 5-step depth, 9-step depth and max_depth=None which means all leaves nodes are pure.
- Output three files for evaluation:
    1) *eval_SeriousDlqin2yrs.csv* (Table 3) provides accuracy scores with *SeriousDlqin2yrs* being the target variable. Because *SeriousDlqin2yrs* is a dummy variable, no discretization function is used. Hence, evaluate it solely.
    2) *eval_Normalization.csv* (Table 4) provides accuracy scores for each variable discretized and categorized by normalization.
    3) *eval_Quartilzation.csv* (Table 5) provides accuracy scores for each variable discretized and categorized in quantile-base. Note that only 4 variables are feasible for this type of discretization: *DebtRatio*, *MonthlyIncome*, *NumberOfOpenCreditLinesAndLoans*, and *age*.
    4) The accuracy score for all variables discretized by normalization reached to 97%-98% whereas that for variables discretized in quantile-base reached only to 60%-70%. Be cautious about the high accuracy rate shown in normalization method because even I used log transformation to deal with the skewness, the skewness of distribution after the transformation can still be high (some log-transformed variables still show left skewness). So high accuracy score in this case may tell us little about what this model has really accomplished.

7. Using the pipeline: (refer to Main_algorithm.py)

- This file import my pipeline-library and use it to solve the problem for this homework.

Table. 1

	count	mean	std	min	25%	50%	75%	max	missing_data_counts
SeriousDlqin2yrs	41016	0.161400429	0.367904377	0	0	0	0	1	0
RevolvingUtilizationOfUnsecuredLines	41016	6.375870039	221.6189498	0	0.034310098	0.189730278	0.6671596 7	22000	0
age	41016	51.68348937	14.74687979	21	41	51	62	109	0
NumberOfTime30-59DaysPastDueNotWorse	41016	0.5892334 7	5.20562764 7	0	0	0	0	98	0
DebtRatio	41016	331.4581373	1296.109695	0	0.176375264	0.3697356 8	0.866470626	106885	0
MonthlyIncome	33042	6578.995733	13446.82593	0	3333	5250	8055.75	1794060	7974
NumberOfOpenCreditLinesAndLoans	41016	8.403476692	5.207323925	0	5	8	11	56	0
NumberOfTimes90DaysLate	41016	0.419592354	5.190382087	0	0	0	0	98	0
NumberRealEstateLoansOrLines	41016	1.008801443	1.153825588	0	0	1	2	32	0
NumberOfTime60-89DaysPastDueNotWorse	41016	0.371586698	5.169641138	0	0	0	0	98	0
NumberOfDependents	39979	0.773230946	1.121269048	0	0	0	1	13	1037

Table. 2

	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
SeriousDlqin2yrs	1	-0.004586161	-0.173727844	0.149333554	-0.013502207	-0.032809646	-0.039897663	0.139609165	-0.010640863	0.121886468	0.065707974
RevolvingUtilizationOfUnsecuredLines	-0.004586161	1	-0.008003426	-0.001999118	0.022250087	0.005831837	-0.014589882	-0.001685767	0.004762918	-0.001413399	0.005342053
age	-0.173727844	-0.008003426	1	-0.068695717	0.038828358	0.048137653	0.159866192	-0.069036148	0.049167595	-0.063622112	-0.211002254
NumberOfTime30-59DaysPastDueNotWorse	0.149333554	-0.001999118	-0.068695717	1	-0.011619699	-0.015223779	-0.070703856	0.984464744	-0.037863365	0.988529725	-0.007839684
DebtRatio	-0.013502207	0.022250087	0.038828358	-0.011619699	1	-0.022987802	0.082791096	-0.014789969	0.177858403	-0.01328969	-0.070558084
MonthlyIncome	-0.0328096	0.005831837	0.04813765	-0.015223779	-0.022987	1	0.107099816	-0.01795368	0.127312823	-0.015336268	0.060528004

	46		3		802						
NumberOfOpenCreditLinesAndLoans	-0.039897663	-0.014589882	0.159866192	-0.070703856	0.082791096	0.107099816	1	-0.098176442	0.442776307	-0.087153635	0.060218201
NumberOfTimes90DaysLate	0.139609165	-0.001685767	-0.069036148	0.984464744	-0.014789969	-0.017953368	-0.098176442	1	-0.054661271	0.992142508	-0.015737468
NumberRealEstateLoansOrLines	-0.010640863	0.004762918	0.049167595	-0.037863365	0.177858403	0.127312823	0.442776307	-0.054661271	1	-0.047995893	0.114879638
NumberOfTime60-89DaysPastDueNotWorse	0.121886468	-0.001413399	-0.063622112	0.988529725	-0.01328969	-0.015336268	-0.087153635	0.992142508	-0.047995893	1	-0.016492604
NumberOfDependents	0.065707974	0.005342053	-0.211002254	-0.007839684	-0.070558084	0.060528004	0.060218201	-0.015737468	0.114879638	-0.016492604	1

Table. 3

SeriousDlqin2yrs	Max_depth	Train_accuracy	Test_accuracy
0	1	0.869910487	0.864851686
1	3	0.872557556	0.866151971
2	5	0.877851694	0.871840715
3	9	0.892201595	0.86875254
4		0.999756191	0.816009752

Table. 4

Variables		Max_depth	Train_accuracy	Test_accuracy
DebtRatio	0	1	0.881926788	0.882324258
	1	3	0.978962767	0.978870378
	2	5	0.980808749	0.979439252
	3	9	0.985092822	0.979032913
	4		0.999791021	0.97220642
MonthlyIncome	0	1	0.982097454	0.981552215
	1	3	0.982167114	0.981470947
	2	5	0.982898541	0.981145876
	3	9	0.985684929	0.978545307
	4		0.999442722	0.968630638
NumberOfDependents	0	1	0.974678695	0.97293783
	1	3	0.974678695	0.97293783
	2	5	0.974922504	0.972531491
	3	9	0.976977465	0.97001219
	4		0.99996517	0.947826087
NumberOfOpenCreditLinesAndLoans	0	1	0.945177806	0.946038196
	1	3	0.959040089	0.960666396
	2	5	0.962209606	0.963104429
	3	9	0.969767685	0.96123527
	4		0.999129254	0.944737911
NumberOfTime30-59DaysPastDueNotWorse	0	1	0.926578663	0.925802519
	1	3	0.929434711	0.928240553
	2	5	0.931420013	0.929622105
	3	9	0.941694821	0.92612759
	4		1	0.889475823
NumberOfTime60-89DaysPastDueNotWorse	0	1	0.981052558	0.982852499

	1	3	0.981052558	0.982852499
	2	5	0.981366027	0.982283625
	3	9	0.984222075	0.979926859
	4		1	0.967655425
NumberOfTimes90DaysLate	0	1	0.964891505	0.963917107
	1	3	0.967677893	0.966842747
	2	5	0.968270001	0.966436408
	3	9	0.97345965	0.962291751
	4		1	0.945225518
NumberRealEstateLoansOrLines	0	1	0.974643865	0.974725721
	1	3	0.974992163	0.973913043
	2	5	0.976385358	0.974725721
	3	9	0.98307269	0.973587972
	4		1	0.963104429
RevolvingUtilizationOfUnsecuredLines	0	1	0.996900143	0.996505486
	1	3	0.996934973	0.996424218
	2	5	0.997074292	0.996261682
	3	9	0.997840549	0.995449004
	4		1	0.9926859
age	0	1	0.969732855	0.97220642
	1	3	0.969732855	0.97220642
	2	5	0.970081154	0.971637546
	3	9	0.972867542	0.968711906
	4		0.997701229	0.94806989

Table. 5

Variables		Max_depth	Train_accuracy	Test_accuracy
DebtRatio	0	1	0.658980878	0.661519707
	1	3	0.805684233	0.801056481
	2	5	0.821392498	0.818691589
	3	9	0.863118665	0.835351483
	4		0.998955104	0.79341731
MonthlyIncome	0	1	0.667723172	0.661032101
	1	3	0.738079482	0.732141406
	2	5	0.764201874	0.763023161
	3	9	0.800355265	0.777732629
	4		0.998014698	0.719138562
NumberOfOpenCreditLinesAndLoans	0	1	0.644665808	0.639902479
	1	3	0.68043607	0.675091426
	2	5	0.698652085	0.688988216
	3	9	0.740064784	0.709061357
	4		0.999860681	0.645428688
RevolvingUtilizationOfUnsecuredLines	0	1	0.612343701	0.610727347
	1	3	0.645362405	0.644372206
	2	5	0.665703041	0.658431532
	3	9	0.729789976	0.699146688
	4		0.999582042	0.646647704
age	0	1	0.607258542	0.602763104
	1	3	0.667305214	0.659813084
	2	5	0.681063007	0.670134092

	3	9	0.712653687	0.670702966
	4		0.997736059	0.609427062