Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Group Number: Team 3

Assignment Title: Group Assignment 1

Course Code: RSM 8413

Instructor Name: Gerhard Trippen

In submitting this **group** work for grading, we confirm:

- That the work is original, and due credit is given to others where appropriate.
- That all members have contributed substantially and proportionally to each group assignment.
- That all members have sufficient familiarity with the entire contents of the group assignment so as to be able to sign off on them as original work.
- Acceptance and acknowledgement that assignments found to be plagiarized in any way will be subject to sanctions under the University's Code of Behaviour on Academic Matters.

Please **check the box and record your student number** below to indicate that you have read and abide by the statements above:

☑ 1004654527
☑ 1006507512
☑ 1002140541
☑ 1005605374
☑ 1006604934
☑ 1006527741

Assignments are to be submitted using Student ID Numbers *only*; do not include your name. Assignments that include names or that do not have the box above checked **will not be graded.**

Please pay attention to Course Outline for specific formatting requirements set by instructors.

EXECUTIVE SUMMARY

Company Background and Problem Identification:

Universal bank is rapidly growing in terms of customer acquisition. However, majority of the newly acquired customers are liability customers, as they only act as depositors with the bank. Thus, the number of customers who act as an asset, by taking out loans with the bank, is relatively small. Universal bank is interested in increasing the number of customers who take out loans with the bank. Moreover, Universal Bank is especially interested in converting its current liability customers into asset customers, while still maintaining them as depositors. The Retail Marketing Department at Universal ran a previous campaign that was aimed at converting liability customers to asset customers, which resulted in a 9.6% conversion rate. The campaign success motivated the Retail Marketing Department to improve their next campaign via better targeted marketing. The improved campaign will utilize a k-Nearest Neighbour (k-NN) classifier algorithm to predict whether a Universal Bank liability customer will accept a loan offer and become an asset customer.

The Universal Bank data set consisted of basic level information on 5,000 customers. After the initial data preparation and exploratory data analysis, which is detailed below, we found very promising results. The k-NN algorithm was able to predict if a customer would accept a personal loan, as part of the improved campaign, with *91.27%* accuracy. This result is a critical indicator of campaign success. The Retail Marketing Department will now be able to accurately identify and target which depositors who are most likely to convert to loan customers. All relevant methodology surrounding the k-NN algorithm is presented below.

DATA PREPARATION

Data Cleaning

Prior to developing the k-NN algorithm, the Universal Bank dataset needed to undergo data preprocessing. The overarching objective of data preprocessing and data cleaning is to ensure that the data is relevant and consistent, and that there are no missing fields. This step is essential as it minimizes the likelihood of ‘garbage in, garbage out,’ or feeding low quality data into the algorithm. When cleaning the data, we observed two major concerns. First, we noticed that there were 53 observations where customers had negative work experience. As defined in the dataset, work experience measures professional experience in years, for a given customer. Therefore, a customer cannot have negative work experience. The least amount of experience a customer can have is zero years. The issue is likely a result of incorrect data entry or a 'fat finger' error. In order to correct this issue, we decided to delete these observations, as they only account for approximately 1% of the total set. Although, more data is almost always better, the difference in sample size is negligible.

Further, we also observed one customer (ID = 385) having ZIP Code 9307, which is only 4 digits. Given the fact that all other customers have a 5 digit United States ZIP Code, we assume that this 4 digit ZIP code is an error. Again, we predict that this is the result of a data entry error. Since there is only one ZIP code error, this observation only represents 0.02% of the total dataset. Similar to error in work experience, we decided to get rid this observation as well.

Data Transformation

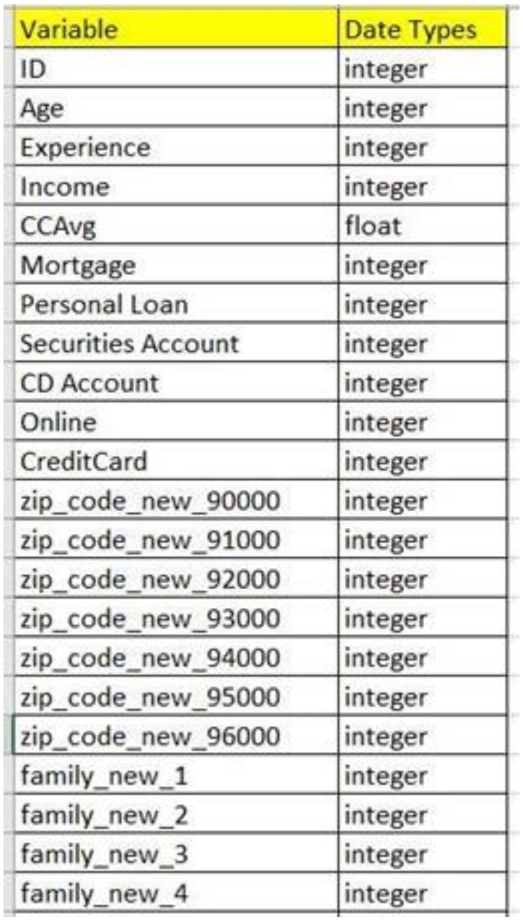
In addition to data cleaning, the data also needed to undergo transformation. This step is to ensure that the data is constructed in a suitable form to be used in the algorithm. Data transformation occurred in two ways. First, we normalized the data in order to standardize the scale of effect that each variable has on the results. This is especially important because k-NN algorithm utilizes a distance measure. The second type of data transformation entailed binning the Zip Code and Family variables, so that they were represented as categorical variables. Lastly, we also transformed the account for the ID variable. The details of the transformations are further explained below.

1. Min/Max Scaler

As the dataset contains both numeric and categorical variables, we preferred the min-max normalization over the default z-score standardization due to the accuracy of results that the former delivers. For consistency reasons we used the same scaler function that we developed in the training dataset for both completely new and unseen test set and the single observation. Since an additional goal of our model is to predict whether the new customer will opt for a Personal loan or not, we need to apply the very same min-max scaler of the training dataset to the single data point, as otherwise it would be impossible to calculate the minimum and maximum values for a single observation.

Data Types

Below we present the variables and their respective data types that were used to develop the k-NN algorithm.



Variable	Date Types
ID	integer
Age	integer
Experience	integer
Income	integer
CCAvg	float
Mortgage	integer
Personal Loan	integer
Securities Account	integer
CD Account	integer
Online	integer
CreditCard	integer
zip_code_new_90000	integer
zip_code_new_91000	integer
zip_code_new_92000	integer
zip_code_new_93000	integer
zip_code_new_94000	integer
zip_code_new_95000	integer
zip_code_new_96000	integer
family_new_1	integer
family_new_2	integer
family_new_3	integer
family_new_4	integer

2. Transforming corresponding to variables

ZIP Code

Since the ZIP code is numeric variable, we needed to apply a data transformation method to transform ZIP code into a binary variable. The primary purpose of this transformation was to make the interpretation of the ZIP code variable more intuitive. The range of the Zip codes is expansive and runs from 90*** to 96***. To achieve this, we created new six binary variables by binning ZIP codes with the same first two digits. All zip codes having the first two digits '90,' were binned together and this process was continued for the entirety of the range. Further, this step was also important in order to properly apply Min/Max scaler. Each binary was separated into a new column as the new predictors (zip_code_new_90000, zip_code_new_91000, ..., zip_code_new_96000). If the customer is in that specific ZIP code area they will be assigned a value of 1 or 0 otherwise. Lastly, we removed the original "ZIP Code" variable from our dataset.

Family

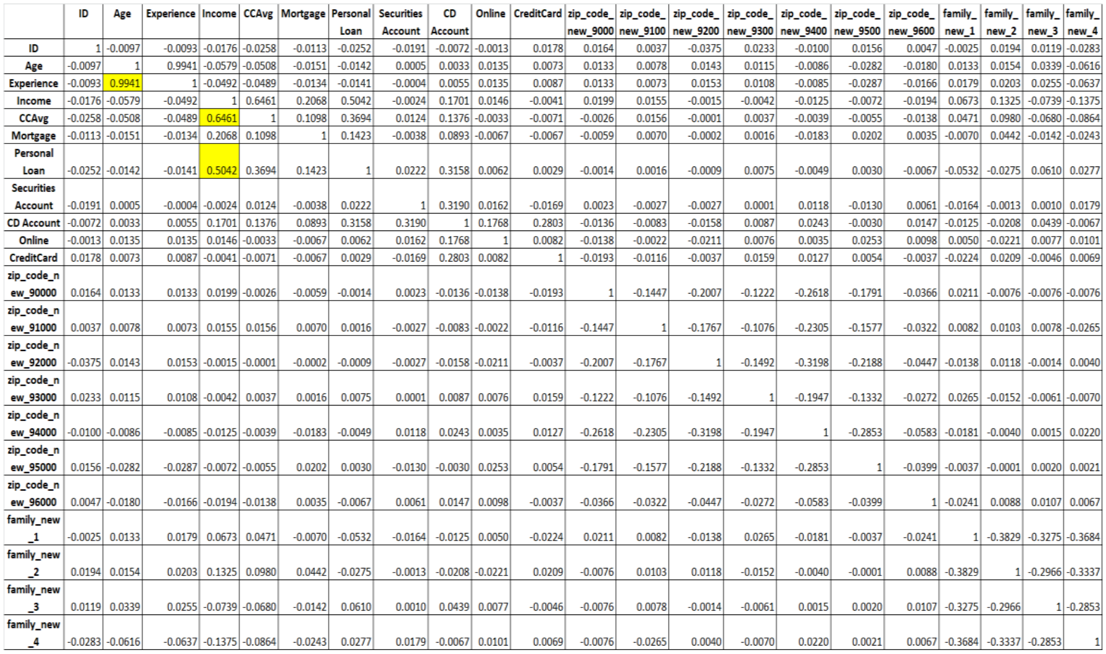
Similarly, the “Family” variable is also numerical ranging from 1 to 4. However, this variable should be transformed into a categorical variable. Normalizing this variable using z-score standardization will result in a float value in range of [0, 1] (e.g. 0.2 family members which seems illogical). As such, we divided them into 4 categorical levels (1, 2, 3, 4) and separated them into 4 different columns as new predictors (family_new_1, family_new_2, family_new_3, family_new_4). Finally, we removed the original “Family” column from our dataset.

ID

Including “ID” variable in our data set is not really helpful when we train our model and make forecasts using our model. ID values are based on the order their information is entered into the system, and thus are not strong indicators of whether the customer will register for the asset program or not. Including ID fields may even cause some spurious relationships between ID and the target variable (i.e. Personal Loan), so we collectively decided to eliminate the “ID” column in the data mining process.

EXPLORATORY DATA ANALYSIS

In order to understand the relationship between the predictors and the target variable (i.e. Personal Loan), we carried out a correlation table between the existing variables, newly created dummy variables, and the target variable. The result is as follows:



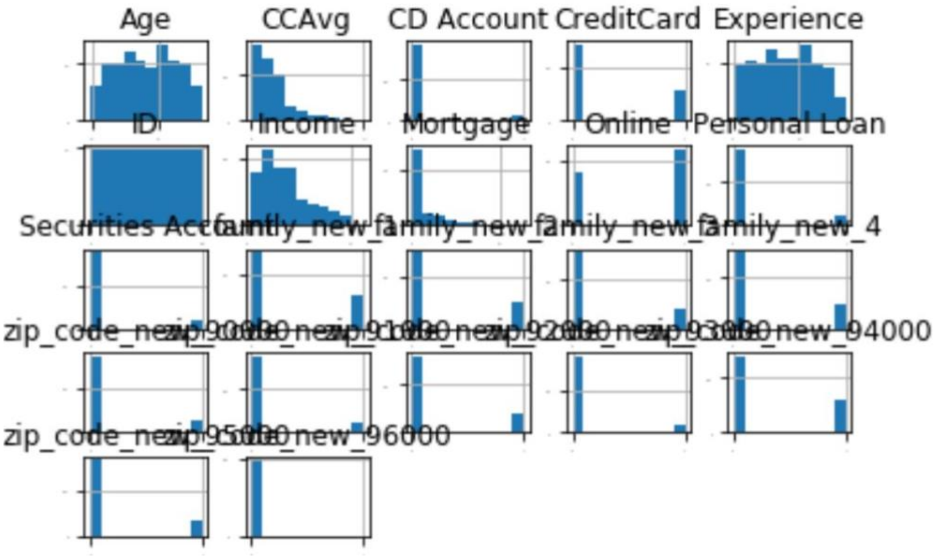
	ID	Age	Experience	Income	CCAvg	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard	zip_code_new_9000	zip_code_new_9100	zip_code_new_9200	zip_code_new_9300	zip_code_new_9400	zip_code_new_9500	zip_code_new_9600	family_new_1	family_new_2	family_new_3	family_new_4
ID	1	-0.0097	-0.0093	-0.0176	-0.0258	-0.0113	-0.0252	-0.0191	-0.0072	-0.0013	0.0178	0.0164	0.0037	-0.0375	0.0233	-0.0100	0.0156	0.0047	-0.0025	0.0194	0.0119	-0.0283
Age	-0.0097	1	0.9941	-0.0579	-0.0508	-0.0151	-0.0142	0.0005	0.0033	0.0135	0.0073	0.0133	0.0078	0.0143	0.0115	-0.0086	-0.0282	-0.0180	0.0133	0.0154	0.0339	-0.0616
Experience	-0.0093	0.9941	1	-0.0492	-0.0489	-0.0134	-0.0141	-0.0004	0.0055	0.0135	0.0087	0.0133	0.0073	0.0153	0.0108	-0.0085	-0.0287	-0.0166	0.0179	0.0203	0.0255	-0.0637
Income	-0.0176	-0.0579	-0.0492	1	0.6461	0.2068	0.5042	-0.0024	0.1701	0.0146	-0.0041	0.0199	0.0155	-0.0015	-0.0042	-0.0125	-0.0072	-0.0194	0.0673	0.1325	-0.0739	-0.1375
CCAvg	-0.0258	-0.0508	-0.0489	0.6461	1	0.1098	0.3694	0.0124	0.1376	-0.0033	-0.0071	-0.0026	0.0156	-0.0001	0.0037	-0.0039	-0.0055	-0.0138	0.0471	0.0980	-0.0680	-0.0864
Mortgage	-0.0113	-0.0151	-0.0134	0.2068	0.1098	1	0.1423	-0.0038	0.0893	-0.0067	-0.0067	-0.0059	0.0070	-0.0002	0.0016	-0.0183	0.0202	0.0035	-0.0070	0.0442	-0.0142	-0.0243
Personal Loan	-0.0252	-0.0142	-0.0141	0.5042	0.3694	0.1423	1	0.0222	0.3158	0.0062	0.0029	-0.0014	0.0016	-0.0009	0.0075	-0.0049	0.0030	-0.0067	-0.0532	-0.0275	0.0610	0.0277
Securities Account	-0.0191	0.0005	-0.0004	-0.0024	0.0124	-0.0038	0.0222	1	0.3190	0.0162	-0.0169	0.0023	-0.0027	-0.0027	0.0001	0.0118	-0.0130	0.0061	-0.0164	-0.0013	0.0010	0.0179
CD Account	-0.0072	0.0033	0.0055	0.1701	0.1376	0.0893	0.3158	0.3190	1	0.1768	0.2803	-0.0136	-0.0083	-0.0158	0.0087	0.0243	-0.0030	0.0147	-0.0125	-0.0208	0.0439	-0.0067
Online	-0.0013	0.0135	0.0135	0.0146	-0.0033	-0.0067	0.0062	0.0162	0.1768	1	0.0082	-0.0138	-0.0022	-0.0211	0.0076	0.0035	0.0253	0.0098	0.0050	-0.0221	0.0077	0.0101
CreditCard	0.0178	0.0073	0.0087	-0.0041	-0.0071	-0.0067	0.0029	-0.0169	0.2803	0.0082	1	-0.0193	-0.0116	-0.0037	0.0159	0.0127	0.0054	-0.0037	-0.0224	0.0209	-0.0046	0.0069
zip_code_new_90000	0.0164	0.0133	0.0133	0.0199	-0.0026	-0.0059	-0.0014	0.0023	-0.0136	-0.0138	-0.0193	1	-0.1447	-0.2007	-0.1222	-0.2618	-0.1791	-0.0366	0.0211	-0.0076	-0.0076	-0.0076
zip_code_new_91000	0.0037	0.0078	0.0073	0.0155	0.0156	0.0070	0.0016	-0.0027	-0.0083	-0.0022	-0.0116	-0.1447	1	-0.1767	-0.1076	-0.2305	-0.1577	-0.0322	0.0082	0.0103	0.0078	-0.0265
zip_code_new_92000	-0.0375	0.0143	0.0153	-0.0015	-0.0001	-0.0002	-0.0009	-0.0027	-0.0158	-0.0211	-0.0037	-0.2007	-0.1767	1	-0.1492	-0.3198	-0.2188	-0.0447	-0.0138	0.0118	-0.0014	0.0040
zip_code_new_93000	0.0233	0.0115	0.0108	-0.0042	0.0037	0.0016	0.0075	0.0001	0.0087	0.0076	0.0159	-0.1222	-0.1076	-0.1492	1	-0.1947	-0.1332	-0.0272	0.0265	-0.0152	-0.0061	-0.0070
zip_code_new_94000	-0.0100	-0.0086	-0.0085	-0.0125	-0.0039	-0.0183	-0.0049	0.0118	0.0243	0.0035	0.0127	-0.2618	-0.2305	-0.3198	-0.1947	1	-0.2853	-0.0583	-0.0181	-0.0040	0.0015	0.0220
zip_code_new_95000	0.0156	-0.0282	-0.0287	-0.0072	-0.0055	0.0202	0.0030	-0.0130	-0.0030	0.0253	0.0054	-0.1791	-0.1577	-0.2188	-0.1332	-0.2853	1	-0.0399	-0.0037	-0.0001	0.0020	0.0021
zip_code_new_96000	0.0047	-0.0180	-0.0166	-0.0194	-0.0138	0.0035	-0.0067	0.0061	0.0147	0.0098	-0.0037	-0.0366	-0.0322	-0.0447	-0.0272	-0.0583	-0.0399	1	-0.0241	0.0088	0.0107	0.0067
family_new_1	-0.0025	0.0133	0.0179	0.0673	0.0471	-0.0070	-0.0532	-0.0164	-0.0125	0.0050	-0.0224	0.0211	0.0082	-0.0138	0.0265	-0.0181	-0.0037	-0.0241	1	-0.3829	-0.3275	-0.3684
family_new_2	0.0194	0.0154	0.0203	0.1325	0.0980	0.0442	-0.0275	-0.0013	-0.0208	-0.0221	0.0209	-0.0076	0.0103	0.0118	-0.0152	-0.0040	-0.0001	0.0088	-0.3829	1	-0.2966	-0.3337
family_new_3	0.0119	0.0339	0.0255	-0.0739	-0.0680	-0.0142	0.0610	0.0010	0.0439	0.0077	-0.0046	-0.0076	0.0078	-0.0014	-0.0061	0.0015	0.0020	0.0107	-0.3275	-0.2966	1	-0.2853
family_new_4	-0.0283	-0.0616	-0.0637	-0.1375	-0.0864	-0.0243	0.0277	0.0179	-0.0067	0.0101	0.0069	-0.0076	-0.0265	0.0040	-0.0070	0.0220	0.0021	0.0067	-0.3684	-0.3337	-0.2853	1

Highlighted in Yellow are some observations of high correlations among the variables in the data set.

Key findings:

- Experience and Age are highly correlated with a correlation of "***0.9941".*** Intuitively, as the higher the customer ages, the higher his/her working experience will be.
- As we observe the correlations between "Personal Loan" and predictors, "Income", "CCAvg" and "CD Account" have the highest correlations respectively. Possible explanations for this are that, as people have more income and dispensable money, they tend to spend more on credit payments and certificate deposits, and thus are more willing to borrow money from banks for their personal purchases.

Also, to gain a deeper understanding on the underlying distribution we plotted the histograms for all the variables within our data set as below.



Key findings:

- The majority of people have reported having only a single member in their family which implies it is just them. Additionally, it can be seen that extremely few people (41) in the dataset belong to the state/province with a starting zip code of “***96”***.

The purpose of our model is to predict whether the customer will accept the loan or not provided the relevant information from the customer. With respect to this new customer, we classify the new customer as 0 (who will not accept the loan), given the k-NN classification using k = 1. Intuitively, it makes sense as well. Just by looking at the important factors, a 40 year old person with $84,000 income (above average), no mortgage, and 2 family members including himself/herself (i.e. the customer is married and has no kids), there is a huge chance that the customer will not be accepting the loan offer. The reason is that as he/she has no mortgage or kids, all of his/her income will be spent just on him/her and their partner. Additionally, there is a possibility that their partner may also be working. Plus, their lives and financial status would be well established at the age of 40.

We split the dataset into training dataset (75%) and test dataset (25%) and then run the k-NN method on the training dataset. For the k-NN method, we first normalize the training dataset using min-max normalization. Then we use the ***same*** parameters of the training dataset (e.g. minimum, maximum) to normalize our test dataset. We use the normalized test dataset to find the average and standard deviation of the mean squared errors of each value of k. The result is as follows:

k (Neighbors)	Average Mean Squared Error (Train Dataset)	Accuracy (Test Dataset)
1	0.0862	0.9111
3	0.0851	0.9127
5	0.0873	0.9119
7	0.0870	0.9078
9	0.0905	0.9087
11	0.0916	0.9038
13	0.0932	0.9022
15	0.0938	0.9006
17	0.0938	0.9006
19	0.0940	0.8998
21	0.0940	0.8998

Out of all k neighbors, we can observe that k = 3 has the highest accuracy rate at 91.27% and lowest mean squared error of 8.5%. As the k value goes up from 3 to 21, the accuracy of the model further decreases and the MSE further increases. Therefore, We choose ***k = 3***, which is neither marginally large or small. This helps us avoid overfitting (k is too small, as noise or outliers may unduly affect classification) and underfitting (k is too large, as large values will tend to smooth out idiosyncratic or obscure data values in the training set) problems in our model.

Below are the steps we implemented to fine tune our model:

- Drop one predictor variable from our dataset: ID variable. Unlike other variables, the ID variable does not contribute to predict whether the customer will register for the personal loan program.

- Incorporate different predictor features for each of categorical variables in the model. For example, we divide Family variable into 4 categorical variables and ZIP Code variable into 7 different bin groups.

- Used the min-max normalization method instead of z-score standardization method.

CLASSIFICATION MATRIX

The classification/confusion matrix on test dataset can be seen in the table below:

Predicted Category				
Actual Category		0	1	Total
	0	1084 (TN)	27 (FP)	1111
	1	81 (FN)	45 (TP)	126
	Total	1165	72	1237

Fields corresponding to ‘1’ in the table above denotes successful conversions. The columns represent the predicted classifications and the rows represent the actual classifications of 1,237 observations in the test data. There are 1,111 people who did not accept to take ‘Personal Loan’ whereas 126 people who did. This gives an ***11.3%*** of successful conversion rate from test data of the campaign. Of 72 records which were predicted as successful conversions by the model, 45 of them were actually converted. However, the model incorrectly classified 27 of these records as unsuccessful conversions. The model gives an accuracy of 91.27% with a sensitivity of 35.7% and a specificity of 97.5%. Therefore, if the bank is interested in obtaining higher sensitivity, in other words higher accuracy of the model to determine actual conversions, then our future task would be to consider different data mining models.

CONCLUSION

The results of our K-NN algorithm will enhance the new campaign at Universal Bank. Our model K-NN will allow the Retail Marketing Department to better predict the customers who will and will not purchase a loan. Ultimately, this greatly improves targeted marketing at Universal Bank.

Final Page

Grade: ____________