

Philip Szymanski

EM 623 Midterm 2016

Phase I – Business Understanding

The purpose of this study is to analyze the dataset `diabetic_data.csv` and create a model to predict under what conditions a patient is readmitted. This will be done using Rattle and Excel. A decision tree will be created using the results of the following data analysis.

Phase II – Data Understanding

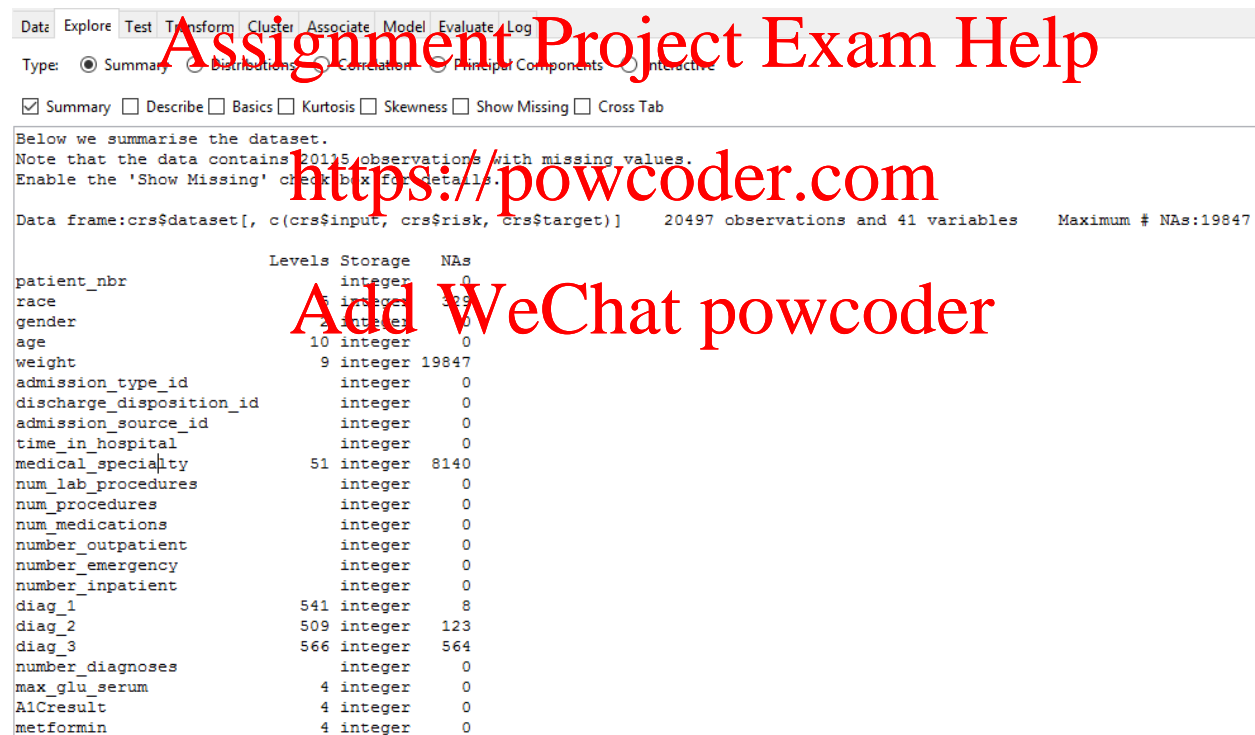
The dataset was first loaded into Rattle. By using the 'Explore' tab in Rattle I was able to determine that the dataset is made up of 41 variables with 20,497 observations. The variables consist of things like race, age, different types of proteins, etc. All somehow relate to diabetes.

According to Rattle 20,115 of the observations have missing values, and there are a total of 19,847 NA's. This means that the dataset will require a lot of cleaning.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Variable	Levels	Storage	NAs
patient_nbr		integer	0
race	5	integer	329
gender	2	integer	0
age	10	integer	0
weight	9	integer	19847
admission_type_id		integer	0
discharge_disposition_id		integer	0
admission_source_id		integer	0
time_in_hospital		integer	0
medical_specialty	51	integer	8140
num_lab_procedures		integer	0
num_procedures		integer	0
num_medications		integer	0
number_outpatient		integer	0
number_emergency		integer	0
number_inpatient		integer	0
diag_1	541	integer	8
diag_2	509	integer	123
diag_3	566	integer	564
number_diagnoses		integer	0
max_glu_serum	4	integer	0
A1Cresult	4	integer	0
metformin	4	integer	0

At the moment the dataset contains too many variables for outlier or distribution analysis to be feasible, it needs to be cleaned and shrunk first.

Phase III – Data Preparation

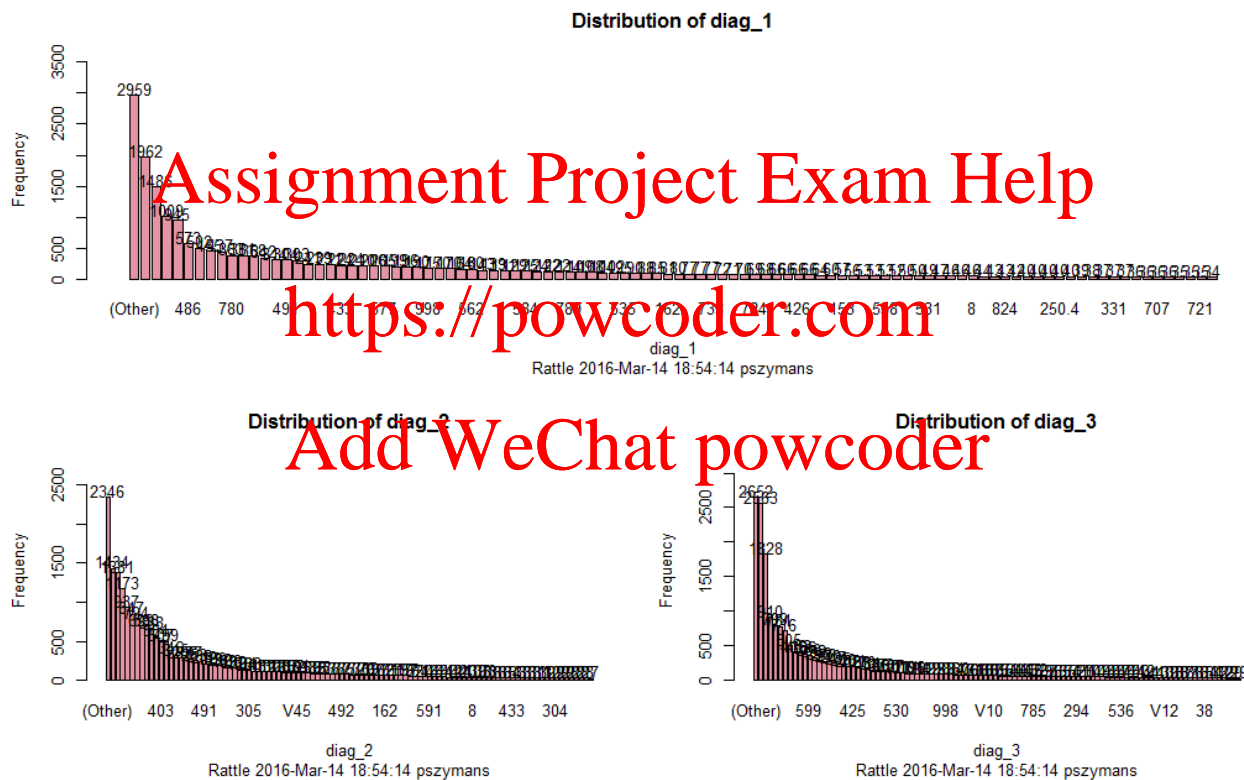
To begin cleaning, variables with more than 50% of their data missing and variables that contained a constant were removed from the dataset with Rattle Software. This was simply done by changing the variable type to 'Ignore' on the 'Data' tab, and then deleting them using the 'Transform'

tab. Next, the dataset was exported into Excel and rows (observations) were analyzed to see if they contained missing data. This was done using the various 'COUNT' statements in Excel. Fortunately none of the rows were missing more than 50% of their data.

The variable 'readmitted' was altered in Excel so that it only contained two values: 'yes' or 'no'. Rows that contained the value 'NA' for the variable were deleted (Only the last row had this problem). This was done using 'IF' statements.

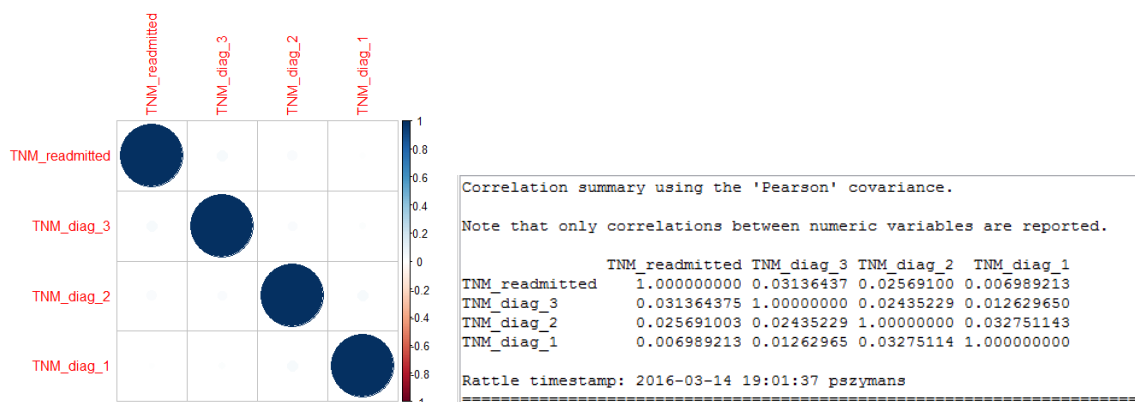
Variables like encounter_id and patient_nbr were deleted because they have no connection with the variable 'readmitted'.

Next variables with a large number of unique entries were analyzed for outliers and for correlation with the variable "readmitted". The variables diag_1-3 all had a large number of unique entries, and the following were their distributions:



These distributions show the existence of many outliers. The variables were then all changed into numerical value so that their relevance could be examined with correlation:

Correlation diabetic_data_saved.csv using Pearson

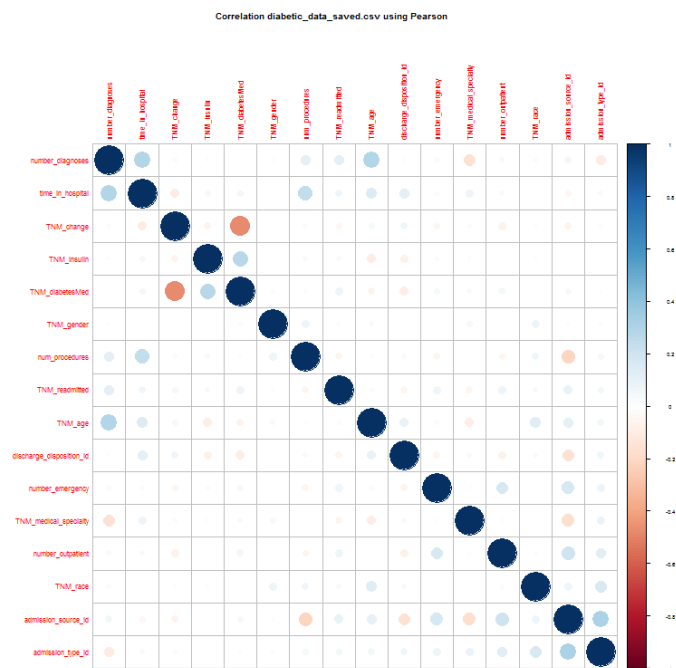


As can be seen there is practically no correlation with the diag_1-3 variables and the variable 'readmitted'. So these variables were deleted for having too many unique categorical entries.

Next all the variables with more than 10 unique entries were analyzed. They were all compared with 'readmitted'. It was found that 'number_inpatient' was quite strongly correlated with 'readmitted' and so this variable was deleted so as to not have adverse effects on the model. The variable time_in_hospital was found to be very correlated to num_lab_procedures and num_medications, which makes sense because the longer you are in the hospital the more medications and procedures you undergo. And so num_lab_procedures and num_medications were deleted.

Next all of the protein or medical term variables were analyzed. They had little to no correlation with the variable 'readmitted'. Over 80% of the data under the variables max_glu_serum and A1Cresult was 'None', and so these variables were deleted. Similarly, over 90% of the data contained in the variables metformin – glyburide.metformin was "No", so these variables were deleted.

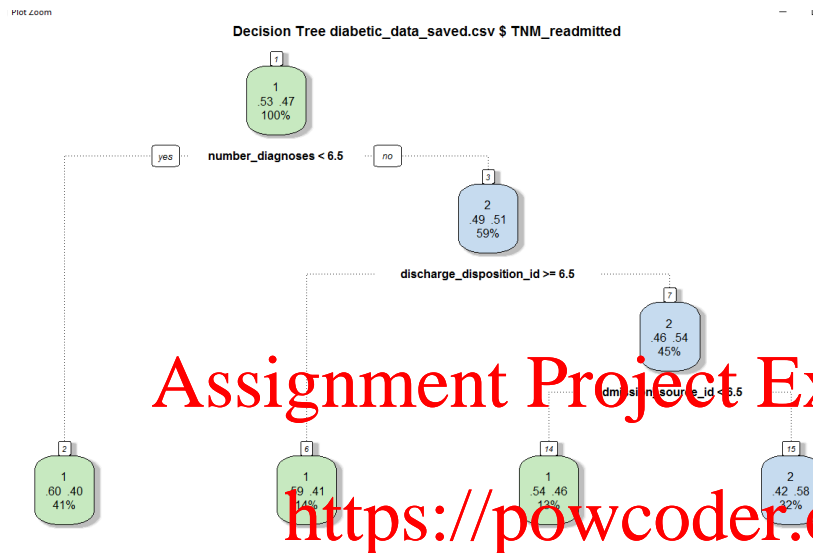
Now that the number of variables was manageable, they were all correlated with one another.



If any two variables were strongly correlated, one of them was removed so as to make the data even more manageable. For example, admission_source_id and admission_type_id were correlated, so one of them was deleted. Medical_specialty was also deleted because it had several thousand missing values and was not strongly correlated to anything.

Phase IV – Modelling

The remaining variables were used to create a decision tree in Rattle. The partition split was set at 70/30. The following decision tree was created:



According to Rattle, if your number of diagnoses is larger than or equal to 6.5, and the discharge disposition id is less than 6.5, and the admission source id is greater than or equal to 6.5 you will be readmitted (2=readmitted).

Tree as rules:

```
Rule number: 15 [TNM_readmitted=2 cover=4590 (32%) prob=0.58]
  number_diagnoses>=6.5
  discharge_disposition_id< 6.5
  admission_source_id>=6.5

Rule number: 14 [TNM_readmitted=1 cover=1896 (13%) prob=0.46]
  number_diagnoses>=6.5
  discharge_disposition_id< 6.5
  admission_source_id< 6.5

Rule number: 6 [TNM_readmitted=1 cover=1958 (14%) prob=0.41]
  number_diagnoses>=6.5
  discharge_disposition_id>=6.5

Rule number: 2 [TNM_readmitted=1 cover=5903 (41%) prob=0.40]
  number_diagnoses< 6.5
```

[1] 7 5 3 1 6 4 2

Phase V – Evaluation

The results of the decision tree created are not great. Whether or not you are readmitted is only a little more accurate than flipping a coin. Unfortunately the inaccuracy of the model is due to time constraints.

The error for this decision tree is about 0.43 according to the error matrixes generated.

```

Error matrix for the Decision Tree model on diabetic_data_saved.csv [test] (counts):

      Predicted
Actual 1 2
1 2345 875
2 1764 1165

Error matrix for the Decision Tree model on diabetic_data_saved.csv [test] (%):

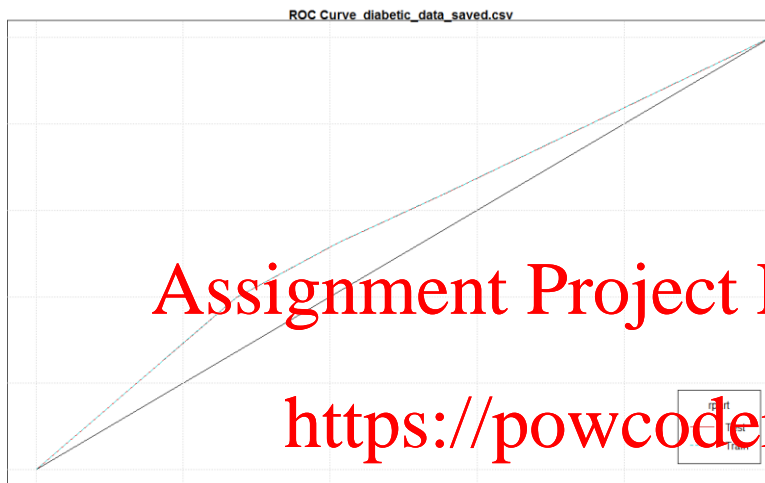
      Predicted
Actual 1 2
1 38 14
2 29 19

Overall error: 0.4291755

Rattle timestamp: 2016-03-14 20:48:11 pszymans
=====

```

And the ROC is as follows:



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Looking back, less of the variables should have been deleted. This could have created a larger and more accurate decision tree. In real life I would go back and prepare the data again in hopes of creating a more accurate tree.

Phase VI – Deployment

The dataset can now be used on new, unsupervised datasets with related variables to predict whether a patient will be readmitted or not.