



CECS 632 – Data Mining
Criteria for grading Project 1 & 2

#	Criteria	Max points	Current Project Student Name: _____ (points)
1	Selection of data set for data mining – how complex, how large, how challenging, real world data?	4	4
2	Understanding and explaining the problem and the goals of data mining project. Is the problem real world or repetition from Internet?	4	3
3	Data preprocessing: explanation of methods applied, and discussion of results obtained. How serious is preprocessing (how noisy are real world data)?	4	3
4	Data mining techniques used in analysis (minimum 3 or more). Explanation of tools used. Tuning the parameters of algorithms used. Presentation of obtained results.	4	2
5	Discussion of results. Are all presented diagrams and tables explained and discussed with enough details. Comparison of methodologies (ROC analysis or similar).	4	2
Total		20	14

Main comments about Project 1 (or 2) report:

_____ > the data is interesting and real world

- The analysis could have been more detailed
- Effects of parameters/model on the results need to be discussed
- Inference on results is important
- Report lacks technical presentation

Data Mining – CECS 632-50

**COMPARING PREDICTIVE ALGORITHMS OF LINEAR REGRESSION, NAÏVE
BAYESIAN AND k-NEAREST NEIGHBORS ON LOUISVILLE CRIME DATA USING
RAPIDMINER**

Wendell Moore – WFM00R01

10/29/2015

Abstract:

In this paper we are comparing three different predictive algorithms - Linear Regression, Naïve Bayesian and k-Nearest Neighbor using the data mining tool RapidMiner. The data used in this project is crime data pulled from the Louisville Metro Government portal - <http://portal.louisvilleky.gov/dataset/crimeadataall-data>. The data is reduced and formatted then it is used to build the predictive models. The predictive values produced by each model will then be compared at the conclusion.

Project:

COMPARING PREDICTIVE ALGORITHMS OF LINEAR REGRESSION, NAÏVE BAYESIAN AND k-NEAREST NEIGHBORS ON LOUISVILLE CRIME DATA USING RAPIDMINER

Predictive analytics is a way for us to prepare for an uncertain future by looking at data from the past. This is true for crime data as well. Building predictive models, identifying trends, can help with any urgent problems law enforcement may have to deal with.

The Louisville Metro Police Department retains historical data on crimes reported in the Louisville Metro area. The data includes details on where the crime occurred, the type of crime as well as dates reported and occurred. An example of the full data set from the portal is on the following page.

Data mining of crime data is very common. Web sites such as www.cityrating.com use local crime statistics to identify safe neighborhoods. Real estate apps such as Zillow and Tulia include crime statistics as well. And crime data is also used by local and national media when reporting on crime trends.

Law enforcement should not be any different than any other business or industry when it comes to mining its data.

INCIDENT_N	DATE_REPOI	DATE_OCCU	CRIME_TYPE	NIBR	UCR_HIERAR	ATT_COMP	LMPD_DIVIS	LMPD_BEAT	PREMISE_TY	BLOCK_ADD	CITY	ZIP_CODE
80-03-610779	2003-04-01 17:00	2003-04-01 02:00	VANDALISM	290	PART II	COMPLETED	2ND DIVISION	223	RESIDENCE / HI	2500 BLOCK W	LOUISVILLE	40210
80-03-610762	2003-04-01 10:00	2003-04-01 10:00	ASSAULT	13B	PART II	COMPLETED	4TH DIVISION	436	RESIDENCE / HI	4600 BLOCK BE	LOUISVILLE	40215
80-03-610910	2003-04-02 14:00	2003-04-02 02:00	VEHICLE BREAK	23F	PART II	COMPLETED	5TH DIVISION	521	PARKING LOT /	400 BLOCK LINE	LOUISVILLE	40206
80-03-610910	2003-04-02 14:00	2003-04-02 02:00	VEHICLE BREAK	23F	PART I	COMPLETED	5TH DIVISION	521	PARKING LOT /	400 BLOCK LINE	LOUISVILLE	40206
80-03-610901	2003-04-02 05:00	2003-04-02 04:00	VANDALISM	290	PART II	COMPLETED	2ND DIVISION	234	RESIDENCE / HI	1700 BLOCK S 3	LOUISVILLE	40211
80-03-610838	2003-04-02 08:00	2003-04-02 03:00	MOTOR VEHICL	240	PART I	COMPLETED	4TH DIVISION	412	HIGHWAY / RO	S 3RD ST / W BL	LOUISVILLE	40208
80-03-707965	2003-04-01 19:00	2003-04-01 18:00	VANDALISM	290	PART II	COMPLETED	8TH DIVISION	812	RESIDENCE / HI	5800 BLOCK BR	LOUISVILLE	40222
80-03-707979	2003-04-02 00:00	2003-03-27 13:00	BURGLARY	220	PART I	COMPLETED	8TH DIVISION	824	RESIDENCE / HI	300 BLOCK STO	LOUISVILLE	40223
80-03-708000	2003-04-02 08:00	2003-04-01 20:00	ASSAULT	13C	PART II	COMPLETED	8TH DIVISION	815	COMMERCIAL /	9200 BLOCK SH	LOUISVILLE	40222
80-03-708029	2003-04-02 15:00	2003-04-01 23:00	BURGLARY	220	PART I	COMPLETED	8TH DIVISION	824	RESIDENCE / HI	1200 BLOCK BL	LOUISVILLE	40299
80-03-708030	2003-04-02 16:00	2003-04-01 15:00	BURGLARY	220	PART I	ATTEMPTED	8TH DIVISION	811	RESIDENCE / HI	10400 BLOCK W	LOUISVILLE	40059
80-03-708041	2003-04-02 21:00	2003-04-01 16:00	VEHICLE BREAK	23F	PART I	COMPLETED	8TH DIVISION	815	GOVERNMENT	9400 BLOCK MI	LOUISVILLE	40223
80-03-610760	2003-04-01 03:00	2003-04-01 02:00	VANDALISM	290	PART II	COMPLETED	6TH DIVISION	611	RESIDENCE / HI	1400 BLOCK NIK	LOUISVILLE	40213
80-03-610945	2003-04-03 00:00	2003-04-02 06:00	VEHICLE BREAK	23F	PART I	COMPLETED	1ST DIVISION	112	HIGHWAY / RO	600 BLOCK S 18	LOUISVILLE	40203
80-03-610945	2003-04-03 00:00	2003-04-02 06:00	VANDALISM	290	PART II	COMPLETED	1ST DIVISION	112	HIGHWAY / RO	600 BLOCK S 18	LOUISVILLE	40203
80-03-610935	2003-04-02 19:00	2003-04-02 05:00	VEHICLE BREAK	23F	PART I	COMPLETED	6TH DIVISION	625	PARKING LOT /	3800 BLOCK KLI	LOUISVILLE	40218
80-03-707955	2003-04-01 03:00	2003-04-01 01:00	VANDALISM	290	PART II	COMPLETED	6TH DIVISION	613	PARKING LOT /	4600 BLOCK ILL	LOUISVILLE	40213
80-03-610949	2003-04-02 22:00	2003-04-02 01:00	MOTOR VEHICL	240	PART I	COMPLETED	1ST DIVISION	135	HIGHWAY / RO	700 BLOCK E LI	LOUISVILLE	40202
80-03-610933	2003-04-02 19:00	2003-04-02 02:00	VEHICLE BREAK	23F	PART I	COMPLETED	5TH DIVISION	512	PARKING LOT /	1800 BLOCK ED	LOUISVILLE	40204
80-03-610930	2003-04-02 18:00	2003-04-02 12:00	VANDALISM	290	PART II	COMPLETED	4TH DIVISION	412	PARKING LOT /	1600 BLOCK S P	LOUISVILLE	40217
80-03-610921	2003-04-02 16:00	2003-04-02 22:00	MOTOR VEHICL	240	PART I	COMPLETED	2ND DIVISION	211	HIGHWAY / RO	600 BLOCK LINE	LOUISVILLE	40211
80-03-610922	2003-04-02 17:00	2003-04-02 20:00	BURGLARY	220	PART I	COMPLETED	4TH DIVISION	424	OTHER / UNKN	3300 BLOCK BO	LOUISVILLE	40215
80-03-708049	2003-04-02 00:00	2003-04-02 22:00	ASSAULT	13B	PART II	COMPLETED	3RD DIVISION	315	RESIDENCE / HI	7000 BLOCK AS	LOUISVILLE	40272
80-03-610932	2003-04-02 18:00	2003-04-02 16:00	ASSAULT	13A	PART I	COMPLETED	4TH DIVISION	411	RESIDENCE / HI	500 BLOCK W C	LOUISVILLE	40203
80-03-610932	2003-04-02 18:00	2003-04-02 16:00	ASSAULT	100	PART II	COMPLETED	4TH DIVISION	411	RESIDENCE / HI	500 BLOCK W C	LOUISVILLE	40203

RapidMiner is the software tool being used in this project. It is an open source, data mining platform developed and maintained by RapidMiner, Inc. and was developed at the University of Dortmund in Germany. It is a GUI-based software where data mining and predictive analytics workflow can be built and deployed.

The software is freely downloadable from www.rapidminer.com. The software contains tutorials and additional training resources can be found online.

For RapidMiner the data will be imported into a tabular form. For each algorithm, the data will be made up of special attributes which are columns that are used for identification and label attributes; these are the attributes to be predicted.

The data for this project, as it was pulled from the Louisville Metro Government portal, is quite detailed, see the example at the top of this page. In order for us not to have to deal with the

“curse of dimensionality”, reduction is necessary. Using software tools such as Access and Excel I was able to make reductions to the data from its original data set.

For this project, we are building and comparing three different predictive models, so I reduced down to a data set that contains only Total counts for crimes reported per LMPD Division, per year, for the month of November of each year. See example on the following page. I compiled 10 years’ worth of data (2005 – 2014) which will be used in the building of the different models, then those models will be used to predict the totals per LMPD Division for November of 2015.

MONTH	YEAR	LMPD_DIVISION	TOTAL
11	2005	1ST DIVISION	977
11	2005	2ND DIVISION	789
11	2005	3RD DIVISION	835
11	2005	4TH DIVISION	1305
11	2005	5TH DIVISION	401
11	2005	6TH DIVISION	813
11	2005	7TH DIVISION	737
11	2005	8TH DIVISION	517
11	2005	METRO LOUISVILLE	243
11	2006	1ST DIVISION	801
11	2006	2ND DIVISION	741
11	2006	3RD DIVISION	837
11	2006	4TH DIVISION	1348
11	2006	5TH DIVISION	496
11	2006	6TH DIVISION	833

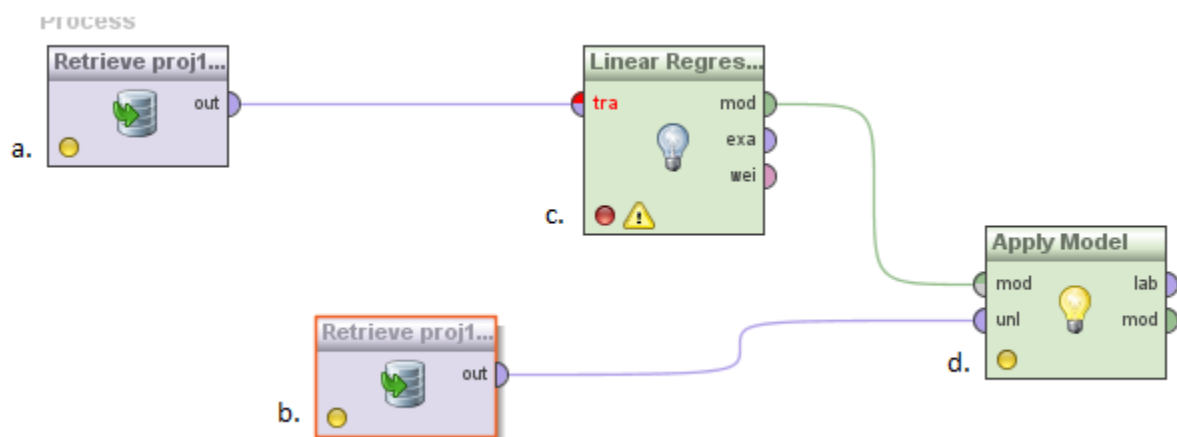
The reduced data set above will be imported into RapidMiner. Per RapidMiner, Month, Year and LMPD_Division will be the special attributes and Total will be the label attribute.

Another data set that is unlabeled is imported as well. This data set will be used for the predicted amounts. See example below.

MONTH	YEAR	LMPD_DIVISION
11	2015	1ST DIVISION
11	2015	2ND DIVISION
11	2015	3RD DIVISION
11	2015	4TH DIVISION
11	2015	5TH DIVISION
11	2015	6TH DIVISION
11	2015	7TH DIVISION
11	2015	8TH DIVISION
11	2015	METRO LOUISVILLE


Both files when imported into RapidMiner and should be saved to the Data folder in the Local Repository. The Local Repository should automatically be set up if you have downloaded RapidMiner. Once all data sets have been imported we will build our models.

Linear Regression



In the Main Process screen of RapidMiner:

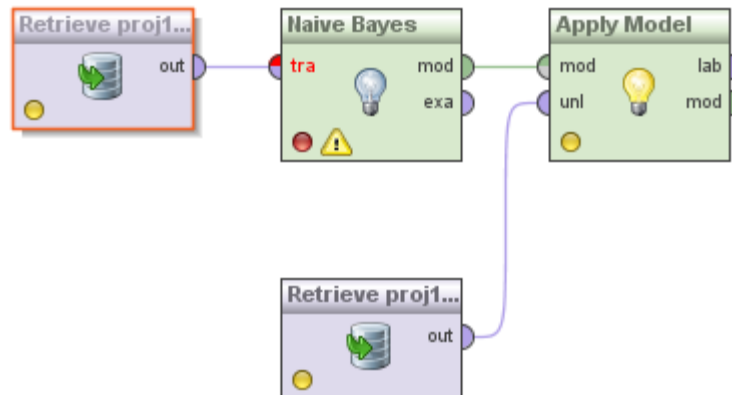
- Retrieve the reduced crime data set that was recently imported.
- Retrieve the unlabeled data set that was recently imported.
- In the Operators section of RapidMiner find Modeling > Classification and Regression > Function Fitting, then select Linear Regression.
- In the Operators section of RapidMiner find Modeling > Model Application, then select Apply Model.

Connect each module to one another as seen in the example above. Press  to run the process. The labeled data from the file in step a is used by the Linear Regression algorithm to build a model and the “Apply Model” module will take the unlabeled data from the file in step b and will produce the output with predictive values. See example below.

Row No.	prediction(T...	MONTH	YEAR	LMPD_DIVISION
1	956.911	11	2015	1ST DIVISION
2	800.811	11	2015	2ND DIVISION
3	1089.711	11	2015	3RD DIVISION
4	1289.911	11	2015	4TH DIVISION
5	490.311	11	2015	5TH DIVISION
6	861.111	11	2015	6TH DIVISION
7	803.011	11	2015	7TH DIVISION
8	552.211	11	2015	8TH DIVISION
9	228.211	11	2015	METRO LOUISVILLE

The remaining builds are similar to the first except for changing the algorithm.

Naïve Bayesian

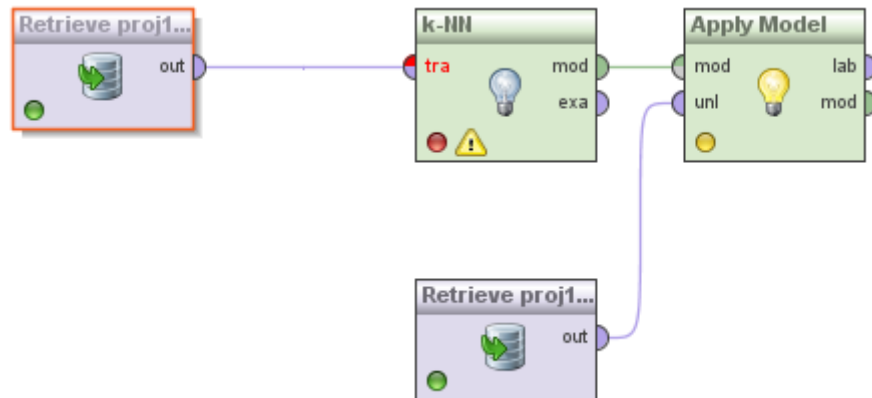


In the Operators section of RapidMiner find Modeling > Classification and Regression > Bayesian Modeling, then select Naïve Bayes.

The predicted values from this model are on the following page.

Row No. ▲	prediction(T...	MONTH	YEAR	LMPD_DIVISION
1	977	11	2015	1ST DIVISION
2	756	11	2015	2ND DIVISION
3	837	11	2015	3RD DIVISION
4	1305	11	2015	4TH DIVISION
5	401	11	2015	5TH DIVISION
6	813	11	2015	6TH DIVISION
7	737	11	2015	7TH DIVISION
8	517	11	2015	8TH DIVISION
9	243	11	2015	METRO LOUISVILLE

k-Nearest Neighbors



In the Operators section of RapidMiner find Modeling > Classification and Regression > Lazy Modeling, then select k-NN.

The predicted values from this model are the following page.

Row No.	prediction(T...	MONTH	YEAR	LMPD_DIVISION
1	839	11	2015	1ST DIVISION
2	836	11	2015	2ND DIVISION
3	1172	11	2015	3RD DIVISION
4	1136	11	2015	4TH DIVISION
5	470	11	2015	5TH DIVISION
6	823	11	2015	6TH DIVISION
7	735	11	2015	7TH DIVISION
8	489	11	2015	8TH DIVISION
9	99	11	2015	METRO LOUISVILLE

Even though we used the same data for input for the models, each algorithm produces different results which is to be expected. To know which of these algorithms works the best with this data requires greater understanding of the data as a whole and a strong subject matter expertise as it relates to law enforcement.

The modeler needs to take these into consideration when developing their models especially if a greater amounts of information is required without adding any complexity to the computation and interpretation of the data.

Appendix:

For further information about RapidMiner as it relates to predictive analytics can be found in –
“Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner” by Vijay Kotu
and Bala Deshpande

Additional information on data mining algorithms are from: “Data Mining: Concepts, Models,
Methods and Algorithms, 2nd edition” by Mehmed Kantardzic.

Online Resources:

<http://portal.louisvilleky.gov/dataset/crimedataall-data>

www.rapidminer.com

<https://hbr.org/2014/09/a-predictive-analytics-primer/>

<http://www.ijcsms.com/> - pdf titled – “Predicting Future Trends in City Crime Using Linear
Regression.

www.ripublication.com/ijaer-spl2/ijaer10n35spl_13.pdf - pdf titled “Classification of Crime Data
Using Rapid Miner”