

DATA MINING IN BUSINESS FIELD

Goktug Cengiz

May 2019

Contents

1	Introduction	3
2	Data Mining for Business	3
3	Open Source Data Mining Tool (KNIME)	5
4	Data set	5
5	Methodology	6
5.1	Using Data Mining Method (CRISP-DM)	6
5.2	Supervised Learning	6
5.2.1	Classification	7
5.3	Unsupervised Learning	7
5.3.1	Clustering	7
5.4	Training Phase	8
6	Implementation with KNIME	8
6.1	Estimated whether the customer will pay the debt or not	9
6.1.1	Business Understanding	9
6.1.2	Data Understanding	9
6.1.3	Data Pre-processing	10
6.1.4	Data Modeling	10
6.1.5	Evaluation	11
6.2	Customer Segmentation	11
6.2.1	Business Understanding	11
6.2.2	Data Understanding	12
6.2.3	Data Pre-processing	12
6.2.4	Data Modeling	13
6.2.5	Evaluation	13
7	Conclusion	16

1 Introduction

Data mining is a field of computer science that refers to extracting useful information from amounts of data, typically huge databases (Data Warehouse). Furthermore, data mining is the process of discovering and analyzing patterns in large data sets involving different methods for retrieve useful information in different areas of our lives such as science, medicine, banking among others. The reason of choosing this process is because it can be used for the field we are involved, business.

In this study we will focus in the importance of data mining and how we can use the methods provided by it in the business field, through a process which we will use one of the best data mining tool (KNIME) for our experiment using two cases ("Customer Segmentation" and "Estimated whether the customer will pay the debt or not") with 2 learning methods (supervised and unsupervised), some of data mining algorithms and a specific business data set.

The objective is to help companies to be one step ahead of the competitors and provide solutions for operating problems and economic issues that can appear day by day.

2 Data Mining for Business

The first step is gathering relevant data critical for the business we are involved, this data can be transactional, non/operational or metadata. The transactional data refers to day by day operations like inventory, sales and cost, non/operational data is normally forecast while metadata is concerned with logical database design. What is really important here is the relationships and patterns among data elements which may increase organizational revenue. For example, companies with a strong consumer focus deal with data mining techniques providing clear pictures of product sold, price, competition and customer demographics, and extracting unknown patterns and possible trends, taking advantage in the market [1].

The second step in data mining is selecting a mechanism producing data model, so there is a need to select a good algorithm for our data. The job of the algorithm is to identify trends in a data set and using the output for parameter definition. In the field of the algorithms for data mining, the most populars are classification algorithms and regression algorithms, which are used to identify the relationship between data elements. The first algorithm is useful when the answer we want to obtain about an specific company falls under a finite set of possible outcomes, meanwhile, the second algorithm is useful for predicting outputs that are continuous, which means that the answer we want obtain is

represented by a quantity that can be flexibly determined based on the inputs of the model rather than being confined to a set of possible labels.

The majority of companies can generate big amounts of data from their work day by day, but this does not mean they have enough budget to expend for Data Mining Tools, and this should not be a reason for them to negate to use the Data Mining process, so here is where there is a need to choose the correct open source tool. The open source data mining tools provide the opportunity to prove the potential of data with minimum costs or even free.

In this section we explain the classifications tasks that we will explore in our experiment and explain how companies can use such tasks to gather information from their data. For our study we focus on the Classification technique because it is the area that has the most use for business environment. First, we explain how classification tasks work in detail, the common uses and naming some use cases for the business field. Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis, and is intended to make the analysis of very large data sets effective.

To begin with a set of training data sets are created with certain set of the attributes or outcomes. Here, the main objective of the algorithm is to mine how that set of attributes reaches its conclusion. We can get too the data about customers, including descriptions of the customers transaction history, for creating segments of customers, for example based on classification but more often based on clustering techniques. Those clusters can build data-driven segments much more optimized than A-B-C segments which even today can often be seen in practice. The assignment of customers to segments is an important prerequisite for further analysis, for example, for selecting customers for debtors and non-debtors can be segmented in two clusters. So, at this point we can realise of the value of the data stored in the digital dispositives of the customers and how important can be in the business field.

It is due to Data Mining techniques that companies are able to use the data to improve their business in any ways, such as estimating whether their customer will pay the debt or not and better profiling customers, and for doing so they need one or more Data Mining tools. That's what we will present in the next section.

3 Open Source Data Mining Tool (KNIME)

Based on the most popular open source data mining tools in the market, we choose to analyze with KNIME. In KNIME [2][3], the user can model workflows, which consist of nodes that process data, transported by means of connections between those nodes. A flow usually starts with a node that reads in data from some data source, but databases can also be queried by special nodes. Imported data is stored in an internal table-based format consisting of columns with a certain (extendable) data type (integer, string, image etc.) and an arbitrary number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes that modify, transform, model, or visualize the data. Modifications can include handling of missing values, filtering of column or rows, oversampling, partitioning of the table into training and test data and many other operators. Following these preparatory steps, predictive models with data mining algorithms such as decision trees, Naive Bayes classifiers or K Nearest Neighbor and clustering algorithms such as k-means are built.

For inspecting the results of an analysis workflow numerous view nodes are available, which display the data or the trained models in diverse ways. In contrast to many other workflow or pipelining tools, nodes in KNIME first process the entire input table before the results are forwarded to successor nodes. Intermediate results can be inspected at any time and new nodes can be inserted and may use already created data without preceding nodes having to be re-executed. One of the node's input are the training (or test) patterns, the output are cluster prototypes.

4 Data set

The chosen data set downloaded from the Kaggle which is called as Loan Data. This data set includes customers who have paid off their loans, who have been past due and put into collection without paying back their loan and interests, and who have paid off only after they were put in collection. The financial product is a bullet loan that customers should pay off all of their loan debt in just one time by the end of the term, instead of an installment schedule. Of course, they could pay off earlier than their pay schedule.

Attribute Information:

- **Loan_id:** A unique loan number assigned to each loan customers
- **Loan_status:** Whether a loan is paid off, in collection, new customer yet to payoff, or paid off after the collection efforts
- **Principal:** Basic principal loan amount at the origination

- **terms:** Can be weekly (7 days), biweekly, and monthly payoff schedule
- **Effective__date:** When the loan got originated and took effects
- **Due__date:** Since it's one-time payoff schedule, each loan has one single due date
- **Paidoff__time:** The actual time a customer pays off the loan
- **Pastdue__days:** How many days a loan has been past due
- **Age, education, gender:** A customer's basic demographic information

5 Methodology

5.1 Using Data Mining Method (CRISP-DM)

To carry out this project, we have followed the CRISP-DM(Cross Industry Standard Process for Data Mining) method, that is a standard process model that describes common approaches used by data mining experts in industry companies. CRISP-DM breaks the process into six phases, these phases can be represented in the scheme in the Figure [1].



Figure 1: CRISP-DM

5.2 Supervised Learning

Supervised learning is typically done in the context of classification, when we want to map input to output labels, or regression, when we want to map input to a continuous output. Common algorithms in supervised learning include Decision Trees and Naive Bayes (We will use these algorithms in Implementation).

In both regression and classification, the goal is to find specific relationships or structure in the input data that allow us to effectively produce correct output data.

5.2.1 Classification

Classification is typically obtained by supervised learning but can also be performed by unsupervised learning, e.g. where the class is not used or unknown as in the Clustering technique. For the test we use algorithms of the following techniques: (1) Decision Tree and (2) Bayesian classifier.

- **Decision Tree:** It is a non-parametric supervised learning method used for classification. Decision tree learn from data to approximate a sine curve. The deeper the tree, the more complex the decision rules and the fitter the model. Moreover, Decision tree constructs classification models in the form of a tree structure. It splits a data set into smaller and smaller subsets while also an associated decision tree is incrementally developed. In addition, A decision node has two or more branches. Finally, Decision trees can handle both categorical and numerical data.
- **Naive Bayes Classifier:** Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our objective is to construct a rule which will permit us to assign future objects to a class, given the vectors of variables describing the future objects.

5.3 Unsupervised Learning

The most common tasks within unsupervised learning are clustering, representation learning, and density estimation. In all of these cases, we wish to learn the inherent structure of our data without using explicitly-provided labels. Some common algorithms include K-Means Clustering (This Algorithm will be used for Customer Segmentation in Implementation Part), principal component analysis, and auto-encoders. Since no labels are provided, there is no specific way to compare model performance in most unsupervised learning methods.

5.3.1 Clustering

- **K-Means:** is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

5.4 Training Phase

Cross-validation is the technique that we used to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. Based on these advantages, we preferred k-fold cross validation instead of using partitioning node and percentage split method. For all of our tests, we used 10 fold cross validation

6 Implementation with KNIME

First, we are reading the loan data set with a *Read CSV* node, then streaming the data to the respective meta nodes. We used meta nodes for both of 2 cases type to keep them organized and easy to manage.

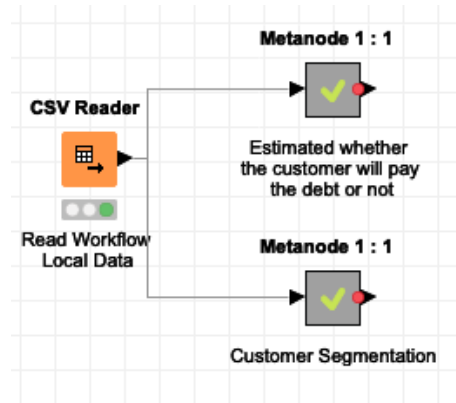


Figure 2: Workflow

Row ID	loan_status	Principal	terms	effective_date	due_date	paid_off_time	past_due_date	age	education	Gender
xqd20166...	PAIDOFF	1000	30	9/8/2016	10/7/2016	9/14/2016 19:31	?	45	High School or Below	male
xqd20168...	PAIDOFF	1000	30	9/8/2016	10/7/2016	10/7/2016 9:00	?	50	Bechalar	female
xqd20160...	PAIDOFF	1000	30	9/8/2016	10/7/2016	9/25/2016 16:58	?	33	Bechalar	female
xqd20160...	PAIDOFF	1000	15	9/8/2016	9/22/2016	9/22/2016 20:00	?	27	college	male
xqd20160...	PAIDOFF	1000	30	9/9/2016	10/8/2016	9/23/2016 21:36	?	28	college	female
xqd20160...	PAIDOFF	300	7	9/9/2016	9/15/2016	9/9/2016 13:45	?	35	Master or Above	male
xqd20160...	PAIDOFF	1000	30	9/9/2016	10/8/2016	10/7/2016 23:07	?	29	college	male
xqd20160...	PAIDOFF	1000	30	9/9/2016	10/8/2016	10/5/2016 20:33	?	36	college	male
xqd20160...	PAIDOFF	1000	30	9/9/2016	10/8/2016	10/8/2016 16:00	?	28	college	male
xqd20160...	PAIDOFF	800	15	9/10/2016	9/24/2016	9/24/2016 13:00	?	26	college	male
xqd20160...	PAIDOFF	300	7	9/10/2016	9/16/2016	9/11/2016 19:11	?	29	college	male
xqd20160...	PAIDOFF	1000	15	9/10/2016	10/9/2016	10/9/2016 16:00	?	39	High School or Below	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/7/2016 23:32	?	26	college	male
xqd20160...	PAIDOFF	900	7	9/10/2016	9/16/2016	9/13/2016 21:57	?	26	college	female
xqd20160...	PAIDOFF	1000	7	9/10/2016	9/16/2016	9/15/2016 14:27	?	27	High School or Below	male
xqd20160...	PAIDOFF	800	15	9/10/2016	9/24/2016	9/24/2016 16:00	?	26	college	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	9/27/2016 14:21	?	40	High School or Below	male
xqd20160...	PAIDOFF	1000	15	9/10/2016	9/24/2016	9/23/2016 18:49	?	32	High School or Below	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/5/2016 22:05	?	32	High School or Below	male
xqd20160...	PAIDOFF	800	30	9/10/2016	10/9/2016	9/23/2016 7:42	?	26	college	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/9/2016 9:00	?	26	college	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/8/2016 17:09	?	43	High School or Below	female
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/9/2016 23:00	?	25	High School or Below	male
xqd20160...	PAIDOFF	1000	15	9/10/2016	9/24/2016	9/24/2016 13:00	?	26	college	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	10/3/2016 12:50	?	26	college	male
xqd20160...	PAIDOFF	1000	30	9/10/2016	10/9/2016	9/29/2016 12:18	?	29	High School or Below	male
xqd20160...	PAIDOFF	800	15	9/10/2016	9/24/2016	9/21/2016 20:16	?	39	Bechalar	male
xqd20170...	PAIDOFF	1000	15	9/10/2016	9/24/2016	9/23/2016 8:21	?	34	Bechalar	male
xqd20160...	PAIDOFF	1000	30	9/11/2016	10/10/2016	9/22/2016 19:17	?	31	college	male
xqd20160...	PAIDOFF	1000	30	9/11/2016	10/10/2016	10/8/2016 17:22	?	22	college	male

Figure 3: Data set

6.1 Estimated whether the customer will pay the debt or not

6.1.1 Business Understanding

Our goal is to make use of data mining algorithms to estimate whether the customer will pay the debt or not.

6.1.2 Data Understanding

In chapter 4 of the article, we mentioned in detail about the data set. It consists of 500 observations and 10 variables. There are 2 types of data type: Numerical and Factor. Factor type loan_status variable will be estimated using variables that affect it. loan_status is in 3 levels. These levels are as follows.

- "COLLECTION": People who have not paid their debts. This is because there are missing values in paid_off_time.
- "COLLECTION_PAIDOFF": People who have paid their debt with delay.
- "PAIDOFF": People who have paid their debts.

6.1.3 Data Pre-processing

In our case it was unnecessary. This is because there are missing values in paid-off, but decision tree can ignore it.

6.1.4 Data Modeling

We chose two classification algorithms Decision Trees and Naive Bayes Classifier. In decision trees, quality measure for the learner node selected as Gini Index and reduced error pruning is active. Starting at the leaves, each node is replaced with its most popular class (selected as loan_status), but only if the prediction accuracy does not decrease. Reduced error pruning has the advantage of simplicity and speed. In Naive Bayes, We used loan_status as the classification column for the Naive Bayes Learner node, with default probability = 0.0001 which is the default value for default probability. Default probability is used when the attribute is nominal and was not seen by the learner or continuous and its probability is smaller than the default probability; which is not true for our case.

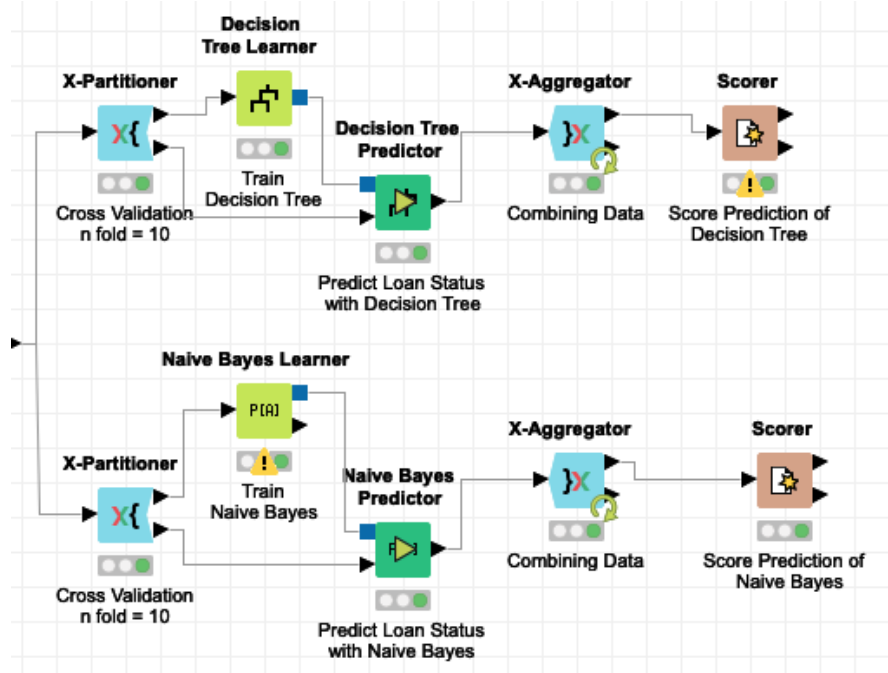
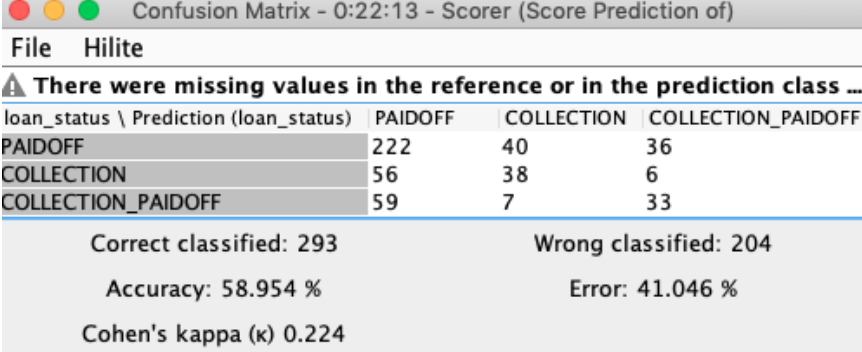


Figure 4: Data Mining Algorithms

6.1.5 Evaluation

We are partitioning the data with *X-Partitioner* node because we are using cross validation method. After executing the respective learner and predictor nodes for each classification methods (details explained in previous sections), we are aggregating the data with *X-Aggregator* and gathered the results with *Scorer* node. Naive Bayes algorithm gave a much better result than decision trees.



Confusion Matrix - 0:22:13 - Scorer (Score Prediction of)

File Hilite

⚠ There were missing values in the reference or in the prediction class ...

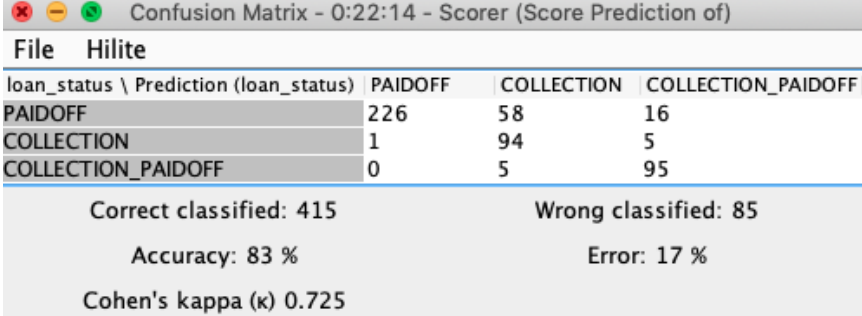
loan_status \ Prediction (loan_status)	PAIDOFF	COLLECTION	COLLECTION_PAIDOFF
PAIDOFF	222	40	36
COLLECTION	56	38	6
COLLECTION_PAIDOFF	59	7	33

Correct classified: 293 Wrong classified: 204

Accuracy: 58.954 % Error: 41.046 %

Cohen's kappa (κ) 0.224

Figure 5: Confusion Matrix for Decision Tree



Confusion Matrix - 0:22:14 - Scorer (Score Prediction of)

File Hilite

loan_status \ Prediction (loan_status)	PAIDOFF	COLLECTION	COLLECTION_PAIDOFF
PAIDOFF	226	58	16
COLLECTION	1	94	5
COLLECTION_PAIDOFF	0	5	95

Correct classified: 415 Wrong classified: 85

Accuracy: 83 % Error: 17 %

Cohen's kappa (κ) 0.725

Figure 6: Confusion Matrix for Naive Bayes

6.2 Customer Segmentation

6.2.1 Business Understanding

Our aim is to divide the customers into two groups as the ones who pay their debts and those who do not pay their debts.

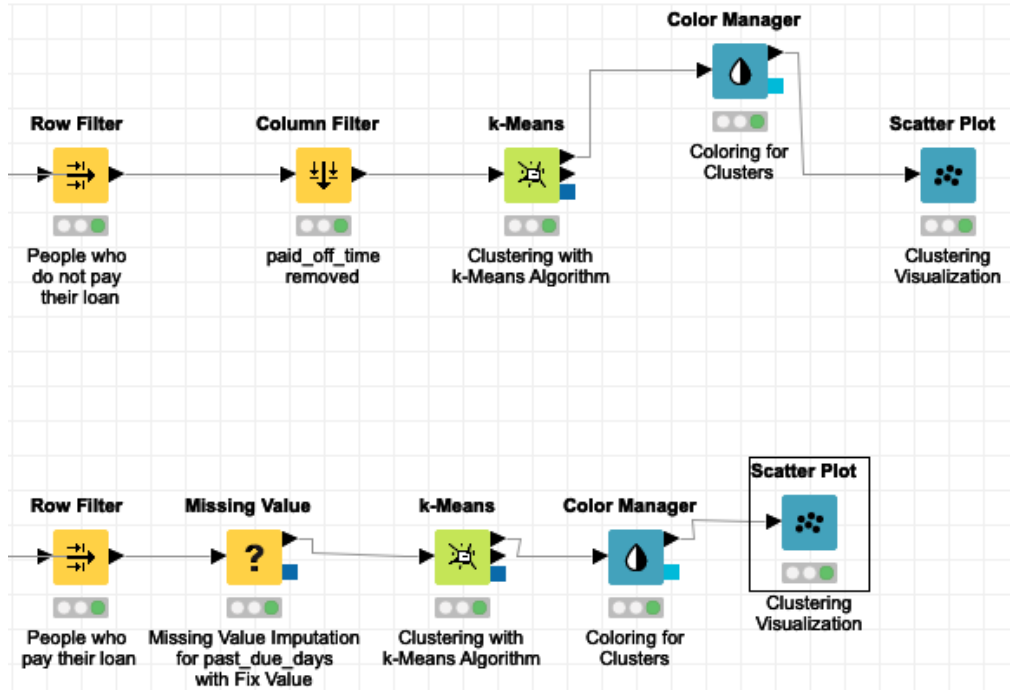


Figure 7: Workflow for Customer Segmentation

6.2.2 Data Understanding

We use the same data set. Details are available in previous sections. Modeling will also use the data of the numerical type. This is because clustering algorithms work with numerical values.

6.2.3 Data Pre-processing

We split the flow of data into two to set the payers among themselves and the non-payers among themselves. Using the row filter node to flow data for those who don't pay the debt to the top we included rows by attribute value (excluded rows by attribute value for other) and used pattern matching with case sensitive match feature. 'paid_off_time' column removed for People who do not pay their loan cluster and missing values imputed for 'past_due_days' with Fix Value.

6.2.4 Data Modeling

After the data was ready, clusters were created with k-means algorithm. The number of clusters is identified as 3 and the max number of iterations is 99.

6.2.5 Evaluation

As you can see below, using the Color Manager and Scatter Plot nodes, clusters are visualized in color. As you can see below, we have divided the clusters who first paid their debt, then those who didn't pay their debt or late payers by age, education and gender.

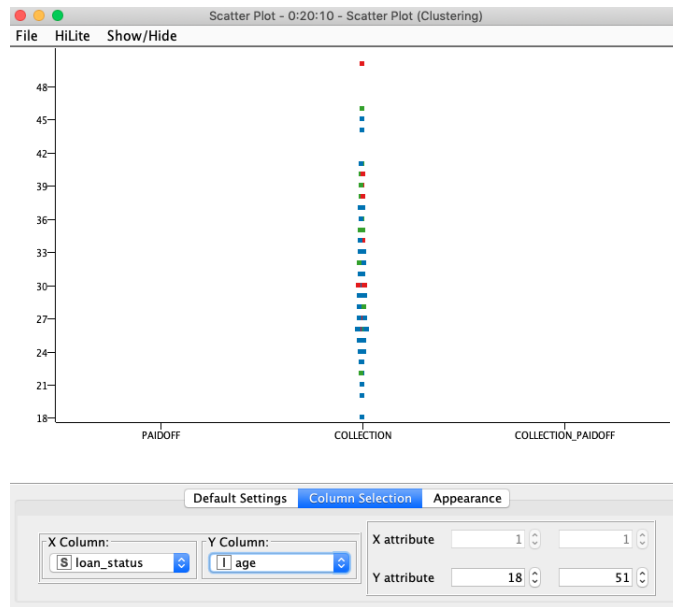


Figure 8: Age for non-Payers

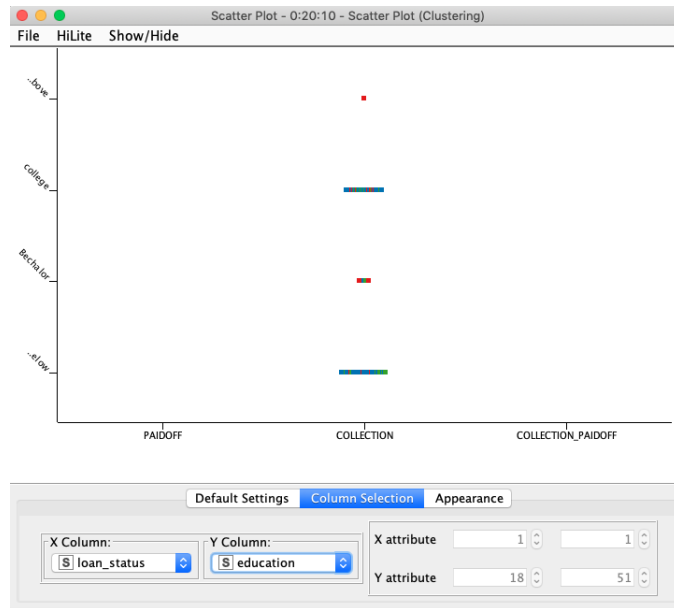


Figure 9: Education Situation for non-Payers

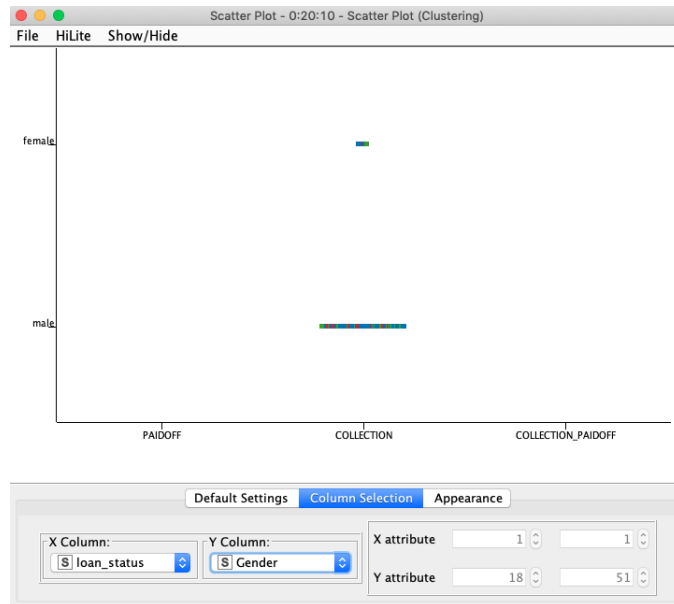


Figure 10: Gender for non-Payers

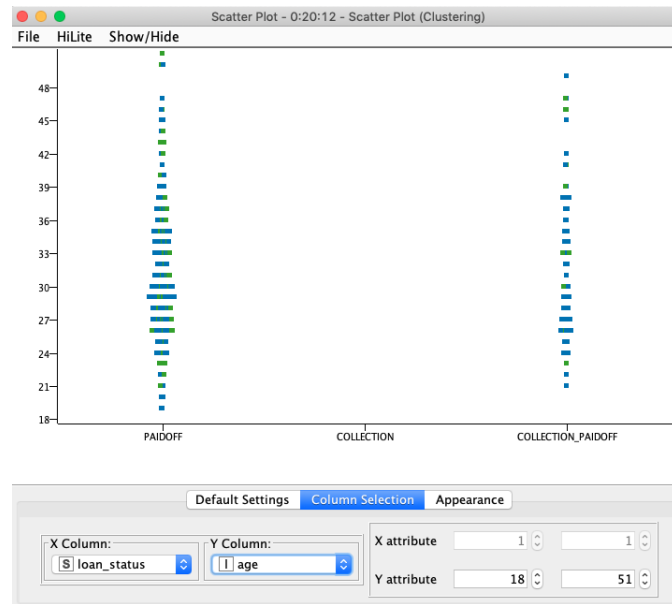


Figure 11: Age for Payers

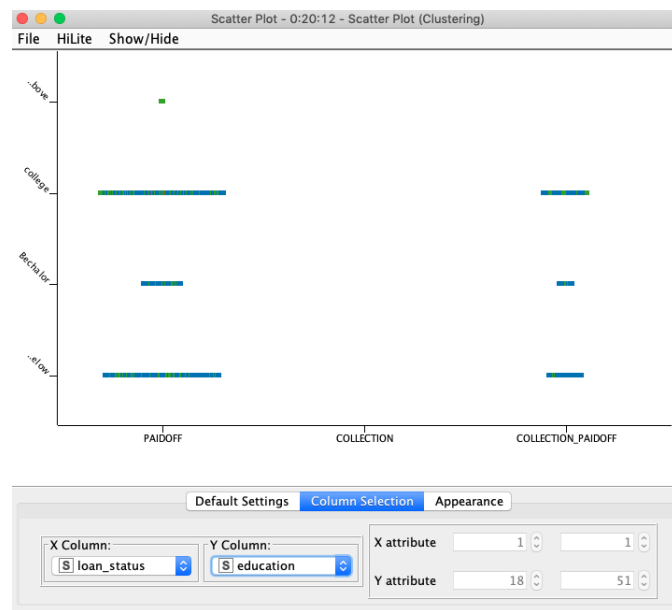


Figure 12: Education Situation for Payers

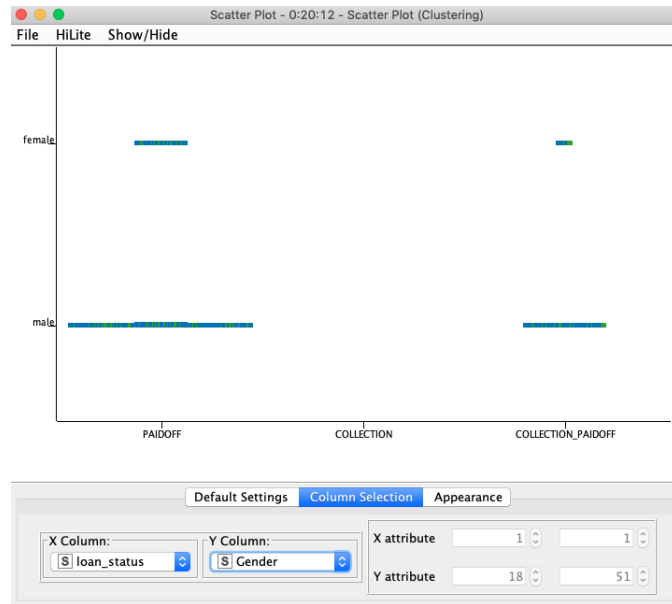


Figure 13: Gender for Payers

7 Conclusion

In this brief case study, we talked about the status of data mining in the business world and what the open source data mining device is. Furthermore, we talked about and applied learning techniques and their algorithms. Finally, we have shown how to apply data mining techniques on a real data set with open source data mining tool.

References

- [1] Almeida, P., Gruenwald, L., & Bernardino, J. (2016). Evaluating Open Source Data Mining Tools for Business. Proceedings of the 5th International Conference on Data Management Technologies and Applications.
- [2] KNIME – Open for Innovation. (n.d.). Retrieved April 26, 2019, from <https://www.knime.com/>
- [3] Knime – EduTech Wiki. (n.d.). Retrieved January April 26, 2019, from <http://edutechwiki.unige.ch/en/Knime>