MET CS-699 DATA MINING FINAL PROJECT REPORT

Team member:

Dongzheng Wang

Weilin Lu

DATA Resource URL:
https://www.kaggle.com/datasets/ytgangster/online-sales-in-usa

Introduction

Data is an essential part of computer science, statistics, data analysis, and business analysis. More and more people are starting to work in the data analysis industry. Therefore, data mining is a must-learn technique for anyone who aspires to become a data analyst. Data mining refers to the process of searching for information hidden in a large amount of data through algorithms. Data mining is generally related to computer science and achieves these goals through several methods such as statistics, online analytical processing, intelligence retrieval, machine learning, expert systems (relying on past rules of thumb), and pattern recognition.

This data mining project finds whether there is a clear relationship between different data (purchase time, purchase quantity, indeed and customer gender) and the payment method chosen by the final customer through transaction data (excluding detailed transaction amount) of a website.

Work distribution

Dongzheng Wang: Find useful data set, clear the data-set. Split the data into 66% training set and 34% test set. Use weka Mining data to come to conclusion. Write how we pre-processing the data and the and what he did during the data mining progress. Mr. Wang did the 60% of the total job.

Weilin Lu:Find useful data set and clear the data-set. Help Mr. Wang split the data into 66% training set and 34% test set. Build the table to show details output of weka. And finish the report paper. Mr. Lu did 40% of the total job.

Detailed description of the datasets

Our data set is the historical purchase records of customers from a shopping website. A total of 26 columns and 500 rows of data. It includes purchase time, status, purchase quantity, price, discount, total, year, month, day, transaction number, name, age, address information and payment method.These data include numeric data and categorical data. All we have to do is to clean the useless data out of the original data, and then mine the correlation between the different data retained and the payment method chosen by the end customer.

Tools to use

The tool we use are ***Chrome*** and ***Weka***. Chrome is a Google's browser. WEKA(Waikato Environment for Knowledge Analysis). It is an open source machine learning and data mining software based on JAVA environment.We use Chrome to find the data-set online and use Weka for data mining to get the confusion matrix

5 Classification Algorithms

Classification is the one of the main technique that we have been learning throughout the semester, it also used for discovering the pattern from known classes. In our real life, we might be facing a data set that contains over hundred attributes, but not every

single attribute is needed for us to complete the mining work. In order to determine which attribute is important, feature selections algorithms are utilized. Instead of processing all the attributes, only relevant attributes are involved in the mining process.

Naïve Bayes:Bayesian classification algorithm is a classification method in statistics, which is a class of algorithms that use the knowledge of probability and statistics to classify.

Logistic Regression:It is a generalized linear regression analysis model, which belongs to supervised learning in machine learning. The model is trained by given n sets of data (training set), and after training is completed, the given set or sets of data (test set) are classified. Each set of data is composed of p indicators

J48:This algorithm is a kind of decision tree. Based on top-to-bottom strategy, recursive divide-and-conquer strategy, select an attribute to place at the root node, generate a branch for each possible attribute value, divide the instance into multiple subsets, each subset corresponds to a branch of the root node , then recursively repeat the process on each branch. Stop when all instances have the same classification.

Random Forest:Random Forest is a classifier that uses multiple trees to train and predict samples.

Bagging:The Bagging algorithm can be combined with other classification and regression algorithms to improve its accuracy and stability, and at the same time reduce the variance of the results to avoid over fitting.

5 attribute selection methods

1.ClassifierAttributeEval:
Evaluates the worth of an attribute by using a user-specified classifier. Here, our attribute will be payment methods.

2.RelifFAttributeEval: Evaluates the worth of an attributes by repeatedly sampling an instance and considering the value of the given attributes for the nearest instance of the same and different class. Can operate on both discrete and continuous class data. We found that relifFattributeEval might be good because we have about 500 rows of dataset.

3.GainRatioAttributeEval: this method measures the significance of attributes with respect to target class on the basis of gain ratio, the follow formula written as GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

4. OneRAttributeEval: class for evaluating attributes individually by using the OneR classifier

5. Correlation Attribute Eval: Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average

Data mining Progress

Data preprocessing is the process of converting raw data into a useful and understandable format. Real world or raw data usually have inconsistent formats, human errors, and may be incomplete. Data pre-processing solves these problems and makes the data set more complete and efficient for data analysis.

This is a critical process that can affect the success of data mining and machine learning projects. It can quickly discover knowledge from the dataset, and ultimately affect the performance of machine learning models. In other words, data pre-processing is to convert data into a form that can be easily processed by the computer. It makes data analysis or visualization easier, and improves the accuracy and speed of machine learning algorithms that train data.

Data pre-processing is very important because a database is a collection of data points. Data points are also called observations, data samples, events, and records. Each sample is described using different characteristics, also known as characteristics or attributes. Data pre-processing is essential to effectively build models with these characteristics. Many problems may arise when collecting data. You may have to aggregate data from different data sources, resulting in data format mismatch, such as integer and floating-point numbers.

The Process of pre-processing our data set.

First, we cleaned our data set, it is a process of cleaning data sets by considering missing values, removing outliers, correcting inconsistent data points, and smoothing noisy data. In essence, the motivation behind data cleaning is to provide complete and accurate samples for machine learning models. The techniques used in data cleansing are specific to the preferences of data scientists and the problems they are trying to solve. We were facing the problem to input our data set into Weka or JMP, because our data set had so much unknown symbol, such as @#& etc. Once we deleted those symbols, we are able to run the test.
We have also engaging a lot of meaningless data is called noise. More precisely, it is the random variance in data with incorrect measurement variables or attribute values. Noise includes repeated or semi repeated data points, data segments that are not valuable for a specific research process, or unnecessary information fields. For example, our goal is to predict whether a person will use Paypal or Visa when they are purchasing. Then, the information about where they live, their hair color, or birth

and height will be irrelevant. The outlier can be considered as insignificant data, although some people think it is a valid data point. Suppose you are training an algorithm to detect the aircraft in the picture. The image data set may contain images of cars that are incorrectly marked as aircraft. This can be considered noise

Data reduction: As the name implies, data reduction is used to reduce the amount of data, thereby reducing the costs associated with data mining or data analysis. The original data set has over thousands of rows, we reduced it with 500 of them are left. I think It provides a compact representation of the data set. Although this step reduces the volume, it maintains the integrity of the original data. When processing big data, this data pre-processing step is particularly important because the amount of data involved will be huge.

Data mining result and evaluation

We data-mined the raw data using five algorithms. After obtaining the confusion matrix of the original data, we trained the data based on five different Attributes and five different algorithms and finally obtained the accuracy and confusion matrix. From the final result comparison, RandomForest based on ClassifierAttributeEval has the highest data accuracy. Its accuracy rate is: 97.2727%

We split the data-set into two parts: 66% as training set, 33% as testing set

5 algorithm only

Naïve Bayes:

Correctly Classified Instances: 90%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.964	0.181	0.871	0.964	0.915	0.800	0.947	0.960	Paypal
	0.819	0.036	0.948	0.819	0.879	0.800	0.946	0.896	Visa
Weighted average	0.900	0.117	0.905	0.900	0.899	0.800	0.947	0.932	

Confusion Matrix

a	b	Classified as
269	10	a=paypal
40	181	b=Visa

Logistic

Correctly Classified Instances: 91%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.918	0.100	0.921	0.918	0.919	0.818	0.938	0.945	Paypal
	0.900	0.08	0.896	0.900	0.898	0.818	0.939	0.904	Visa

Weighted average	0.910	0.092	0.910	0.910	0.910	0.818	0.939	0.927	

Confusion Matrix

a	b	Classified as
256	23	a=paypal
22	199	b=Visa

J48

Correctly Classified Instances: 94%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.982	0.113	0.916	0.982	0.948	0.880	0.981	0.982	Paypal
	0.887	0.018	0.975	0.887	0.929	0.880	0.981	0.969	Visa
Weighted average	0.940	0.071	0.942	0.940	0.940	0.880	0.981	0.976	

Confusion Matrix

a	b	Classified as
274	5	a=paypal
25	196	b=Visa

Random Forest

Correctly Classified Instances: 96.4%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.971	0.045	0.964	0.971	0.968	0.927	0.989	0.989	Paypal
	0.955	0.029	0.963	0.955	0.959	0.927	0.989	0.989	Visa
Weighted average	0.9964	0.038	0.964	0.964	0.964	0.927	0.989	0.989	

Confusion Matrix

a	b	Classified as
271	8	a=paypal
10	211	b=Visa

Bagging

Correctly Classified Instances: 94.8%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.953	0.059	0.953	0.953	0.953	0.895	0.989	0.991	Paypal
	0.941	0.047	0.941	0.941	0.941	0.895	0.989	0.987	Visa

Weighted average	0.948	0.053	0.948	0.948	0.948	0.895	0.989	0.989	

Confusion Matrix

a	b	Classified as
266	13	a=paypal
13	208	b=Visa

Attributes selected :

1) ClassifierAttributeEval:

a) Naïve Bayes

Correctly Classified Instances: 87.5758%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.973	0.247	0.833	0.973	0.879	0.757	0.945	0.958	Paypal
	0.953	0.027	0.957	0.753	0.843	0.757	0.944	0.897	Visa
Weighted average	0.876	0.150	0.887	0.873	0.873	0.757	0.945	0.931	

Confusion Matrix

a	b	Classified as
179	5	a=paypal
36	110	b=Visa

b) Logistics

Correctly Classified Instances: 87.5758%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.875	0.164	0.870	0.875	0.873	0.711	0.905	0.889	Paypal
	0.836	0.125	0.841	0.836	0.838	0.711	0.903	0.872	Visa
Weighted average	0.858	0.147	0.857	0.858	0.858	0.711	0.904	0.882	

Confusion Matrix

a	b	Classified as
161	23	a=paypal
24	122	b=Visa

c) J48

Correctly Classified Instances: 95.4545%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class

	0.978	0.075	0.942	0.978	0.960	0.908	0.984	0.978	Paypal
	0.925	0.022	0.971	0.925	0.947	0.908	0.984	0.978	Visa
Weighted average	0.955	0.052	0.955	0.955	0.954	0.908	0.954	0.978	

Confusion Matrix

a	b	Classified as
180	4	a=paypal
11	135	b=Visa

d) RandomForest:

Correctly Classified Instances: 97.2727%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.984	0.041	0.968	0.984	0.976	0.945	0.995	0.995	Paypal
	0.959	0.016	0.979	0.959	0.969	0.945	0.995	0.995	Visa
Weighted average	0.973	0.303	0.973	0.973	0.973	0.945	0.995	0.995	

Confusion Matrix

a	b	Classified as
181	3	a=paypal
6	140	b=Visa

e) Bagging:

Correctly Classified Instances: 95.1515%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.967	0.068	0.947	0.967	0.957	0.902	0.989	0.991	Paypal
	0.932	0.033	0.958	0.932	0.944	0.902	0.989	0.989	Visa
Weighted average	0.952	0.053	0.952	0.952	0.951	0.902	0.989	0.990	

Confusion Matrix

a	b	Classified as
178	6	a=paypal
10	136	b=Visa

2) RelifFAttributeEval

a) Naïve Bayes

Correctly Classified Instances: 86.3636%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class

	0.832	0.096	0.916	0.832	0.872	0.731	0.955	0.957	Paypal
	0.904	0.168	0.810	0.904	0.854	0.731	0.955	0.957	Visa
Weighted average	0.864	0.128	0.869	0.864	0.864	0.731	0.955	0.957	

Confusion Matrix

a	b	Classified as
153	31	a=paypal
14	132	b=Visa

b) Logistics

Correctly Classified Instances: 92.4242%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.929	0.082	0.934	0.929	0.932	0.847	0.951	0.964	Paypal
	0.918	0.071	0.912	0.918	0.915	0.847	0.951	0.900	Visa
Weighted average	0.924	0.077	0.924	0.924	0.924	0.847	0.951	0.936	

Confusion Matrix

a	b	Classified as
171	13	a=paypal
12	134	b=Visa

c) J48

Correctly Classified Instances: 93.9394%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.951	0.075	0.941	0.951	0.946	0.877	0.981	0.981	Paypal
	0.925	0.049	0.938	0.925	0.931	0.877	0.981	0.971	Visa
Weighted average	0.939	0.064	0.939	0.939	0.939	0.877	0.981	0.976	

Confusion Matrix

a	b	Classified as
175	9	a=paypal
11	135	b=Visa

d) Randomforest

Correctly Classified Instances: 96.0606%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class

	0.978	0.44	0.952	0.978	0.965	0.920	0.990	0.992	Paypal
	0.938	0.022	0.972	0.938	0.955	0.920	0.990	0.983	Visa
Weighted average	0.961	0.062	0.961	0.961	0.961	0.920	0.990	0.988	

Confusion Matrix

a	b	Classified as
180	4	a=paypal
9	137	b=Visa

e) Bagging

Correctly Classified Instances: 94.8485%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.062	0.951	0.957	0.954	0.896	0.988	0.990	Paypal
	0.938	0.043	0.945	0.938	0.942	0.896	0.988	0.987	Visa
Weighted average	0.948	0.054	0.948	0.948	0.948	0.896	0.988	0.989	

Confusion Matrix

a	b	Classified as
176	8	a=paypal
9	137	b=Visa

3) OneRAttributeEval

a) Naïve Bayes

Correctly Classified Instances: 90%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.902	0.103	0.917	0.902	0.910	0.798	0.954	0.953	Paypal
	0.897	0.098	0.879	0.897	0.88	0.798	0.954	0.957	Visa
Weighted average	0.900	0.101	0.900	0.900	0.900	0.798	0.954	0.955	

Confusion Matrix

a	b	Classified as
166	18	a=paypal
15	131	b=Visa

b) Logistics

Correctly Classified Instances: 92.4242%

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.935	0.089	0.930	0.935	0.932	0.846	0.964	0.969	Paypal

	0.911	0.065	0.917	0.911	0.914	0.846	0.964	0.959	Visa
Weighted average	0.924	0.078	0.924	0.924	0.924	0.846	0.964	0.964	

Detailed Accuracy By Class

Confusion Matrix

a	b	Classified as
172	12	a=paypal
13	133	b=Visa

c) J48

Correctly Classified Instances: 92.4242%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.935	0.089	0.930	0.935	0.932	0.846	0.964	0.969	Paypal
	0.911	0.065	0.917	0.911	0.914	0.846	0.964	0.959	Visa
Weighted average	0.924	0.079	0.924	0.924	0.924	0.846	0.964	0.964	

Confusion Matrix

a	b	Classified as
172	12	a=paypal
13	133	b=Visa

d) Randomforest

Correctly Classified Instances: 94.5455%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.068	0.946	0.957	0.951	0.889	0.987	0.990	Paypal
	0.932	0.043	0.944	0.932	0.938	0.889	0.987	0.979	Visa
Weighted average	0.945	0.057	0.945	0.945	0.945	0.889	0.987	0.985	

Confusion Matrix

a	b	Classified as
176	8	a=paypal
10	136	b=Visa

e) Bagging

Correctly Classified Instances: 94.5455%

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.068	0.946	0.957	0.951	0.889	0.987	0.990	Paypal
	0.932	0.043	0.944	0.932	0.938	0.889	0.987	0.979	Visa

Weighted average	0.945	0.057	0.945	0.945	0.945	0.889	0.987	0.985	

Detailed Accuracy By Class

Confusion Matrix

a	b	Classified as
176	8	a=paypal
10	136	b=Visa

4) CorrelationAttributeEval

a) Naïve Bayes

Correctly Classified Instances: 89.697%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.908	0.116	0.908	0.908	0.908	0.791	0.953	0.955	Paypal
	0.887	0.092	0.884	0.884	0.884	0.791	0.953	0.956	Visa
Weighted average	0.897	0.106	0.897	0.897	0.897	0.791	0.953	0.955	

Confusion Matrix

a	b	Classified as
167	17	a=paypal
17	129	b=Visa

b) Logistics

Correctly Classified Instances: 91.5152%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.918	0.089	0.929	0.918	0.923	0.828	0.935	0.959	Paypal
	0.911	0.082	0.899	0.911	0.905	0.828	0.935	0.845	Visa
Weighted average	0.915	0.086	0.915	0.915	0.915	0.828	0.935	0.909	

Confusion Matrix

a	b	Classified as
169	15	a=paypal
13	133	b=Visa

c) J48

Correctly Classified Instances: 94.2424%

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.075	0.941	0..957	0.949	0.883	0.982	0.981	Paypal
	0.925	0.043	0.944	0.925	0.934	0.883	0.982	0.972	Visa

Weighted average	0.942	0.061	0.942	0.9412	0.942	0.883	0.982	0.977	

Detailed Accuracy By Class

Confusion Matrix

a	b	Classified as
176	8	a=paypal
11	135	b=Visa

d) Randomforest

Correctly Classified Instances: 95.4545%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.048	0.962	0.957	0.959	0.908	0.990	0.993	Paypal
	0.952	0.043	0.946	0.952	0.949	0.908	0.990	0.983	Visa
Weighted average	0.955	0.9046	0.955	0.955	0.955	0.908	0.990	0.989	

Confusion Matrix

a	b	Classified as
176	8	a=paypal
7	139	b=Visa

e) Bagging

Correctly Classified Instances: 94.8485%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.062	0.951	0.957	0.954	0.896	0.988	0.990	Paypal
	0.938	0.043	0.945	0.938	0.942	0.896	0.988	0.987	Visa
Weighted average	0.948	0.054	0.948	0.948	0.948	0.896	0.988	0.989	

Confusion Matrix

a	b	Classified as
176	8	a=paypal
9	137	b=Visa

5) GainRatioAttributeEval

a) Naïve Bayes

Correctly Classified Instances: 83.0303%

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.902	0.260	0.841	0.902	0.856	0.656	0.912	0.926	Paypal
	0.740	0.098	0.857	0.740	0.794	0.656	0.912	0.899	Visa

Weighted average	0.830	0.188	0.833	0.830	0.828	0.656	0.912	0.914	

Detailed Accuracy By Class

Confusion Matrix

a	b	Classified as
166	18	a=paypal
38	108	b=Visa

b) Logistics

Correctly Classified Instances: 86.0606%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.870	0.151	0.879	0.870	0.874	0.718	0.924	0.939	Paypal
	0.849	0.130	0.838	0.849	0.844	0.718	0.924	0.855	Visa
Weighted average	0.861	0.142	0.861	0.861	0.861	0.718	0.924	0.902	

Confusion Matrix

a	b	Classified as
160	24	a=paypal
22	124	b=Visa

c) J48

Correctly Classified Instances: 91.2121%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.918	0.096	0.923	0.918	0.921	0.822	0.967	0.966	Paypal
	0.904	0.082	0.898	0.904	0.901	0.822	0.967	0.954	Visa
Weighted average	0.912	0.090	0.912	0.912	0.912	0.822	0.967	0.961	

Confusion Matrix

a	b	Classified as
169	15	a=paypal
14	132	b=Visa

d) Randomforest

Correctly Classified Instances: 93.6364%

Detailed Accuracy By Class

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.957	0.089	0.931	0.957	0.944	0.871	0.985	0.988	Paypal
	0.911	0.043	0.943	0.911	0.927	0.871	0.985	0.983	Visa
Weighted average	0.936	0.069	0.937	0.936	0.936	0.871	0.985	0.986	

Confusion Matrix

a	b	Classified as
176	8	a=paypal
13	133	b=Visa

e) Bagging

Correctly Classified Instances: 90.9091%

Detailed Accuracy By Class

Confusion Matrix

	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	class
	0.929	0.116	0.910	0.929	0.919	0.815	0.974	0.975	Paypal
	0.884	0.071	0.908	0.884	0.896	0.815	0.974	0.972	Visa
Weighted average	0.909	0.096	0.909	0.909	0.909	0.815	0.974	0.975	

a	b	Classified as
171	13	a=paypal
17	129	b=Visa

Conclusion

Summing the above data and accuracy, it can be found that Random Forest based on ClassifierAttributeEval has the highest data accuracy. We can also conclude that even with the same Attribute, using different algorithms can have a huge impact on the results. Therefore, when mining data, it is necessary to conduct multiple experiments to ensure that the best data model is obtained, so as to obtain the most correct data results.