**Project 2 – Decision Trees, Linear Regression, Model Trees, Regression Trees**

**CS548 Knowledge Discovery and Data Mining - Fall 2016**

**Prof. Carolina Ruiz**

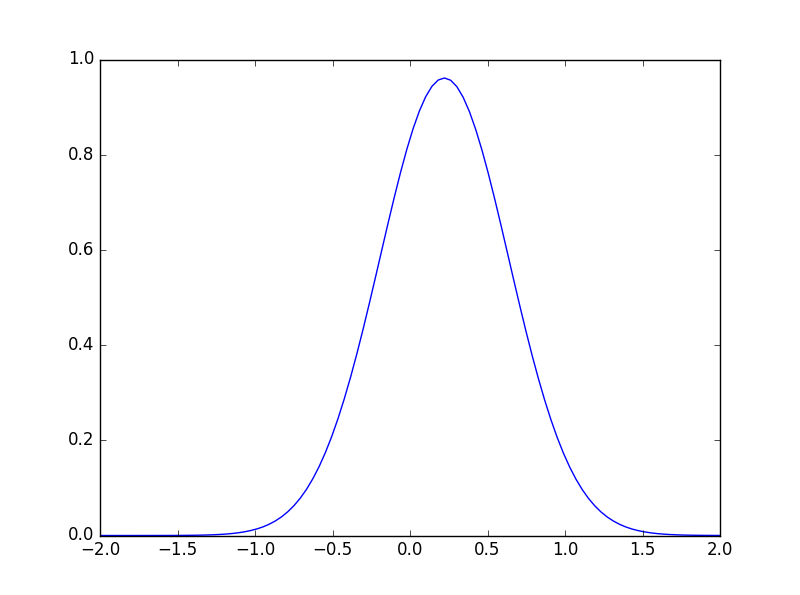
**Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)**

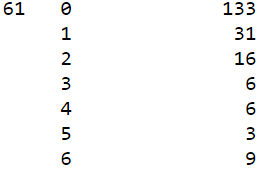
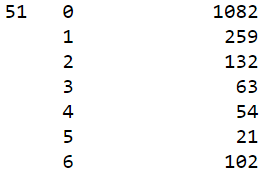
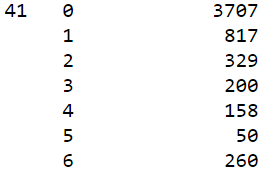
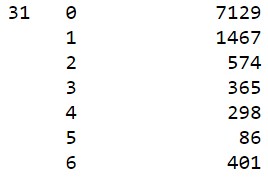
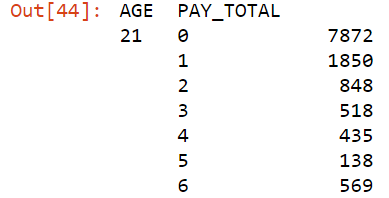
**[05 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)**

The dataset portrays information of a bank’s credit card clients along with their marital status, their balance limit, education, when they paid respective bill amounts. The data is huge with 30000 instances, 23 predictors and a response variable. Skimming through the dataset we found there were no missing values. An error in the naming of variable Pay\_0, it was changed to Pay\_1. Also, certain numeric variables which should be considered as nominal were converted to nominal. These include: SEX, EDUCATION, MARRIAGE, PAY\_1 to Pay\_6 & default payment next month. In the education variable, the values ‘0,4,5,6’ are combined to one value as value ‘4’.

**[10 points] Data Exploration: (e.g., comments on interesting or salient aspects of the dataset, visualizations, correlation, issues with the data, …)**

An interesting observation is that all the attributes have numeric values as the set of values/range for that particular attribute and some of these are converted to nominal as per requirement. This dataset is of the form (n>p) where n is the number of instances and p is the number of attributes.

 The skewness of the target variable is 1.3435 which indicates that it is skewed right. A skewing to the right conveys that there are more ‘0’ i.e. non-defaulters than defaulters and the mean is towards the right. Also, the data has negative kurtosis with a value of ‘-0.19501’. A negative kurtosis represents a flatter value than the regular Gaussian distribution.

****

**[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, necessary dimensionality reduction, …)**

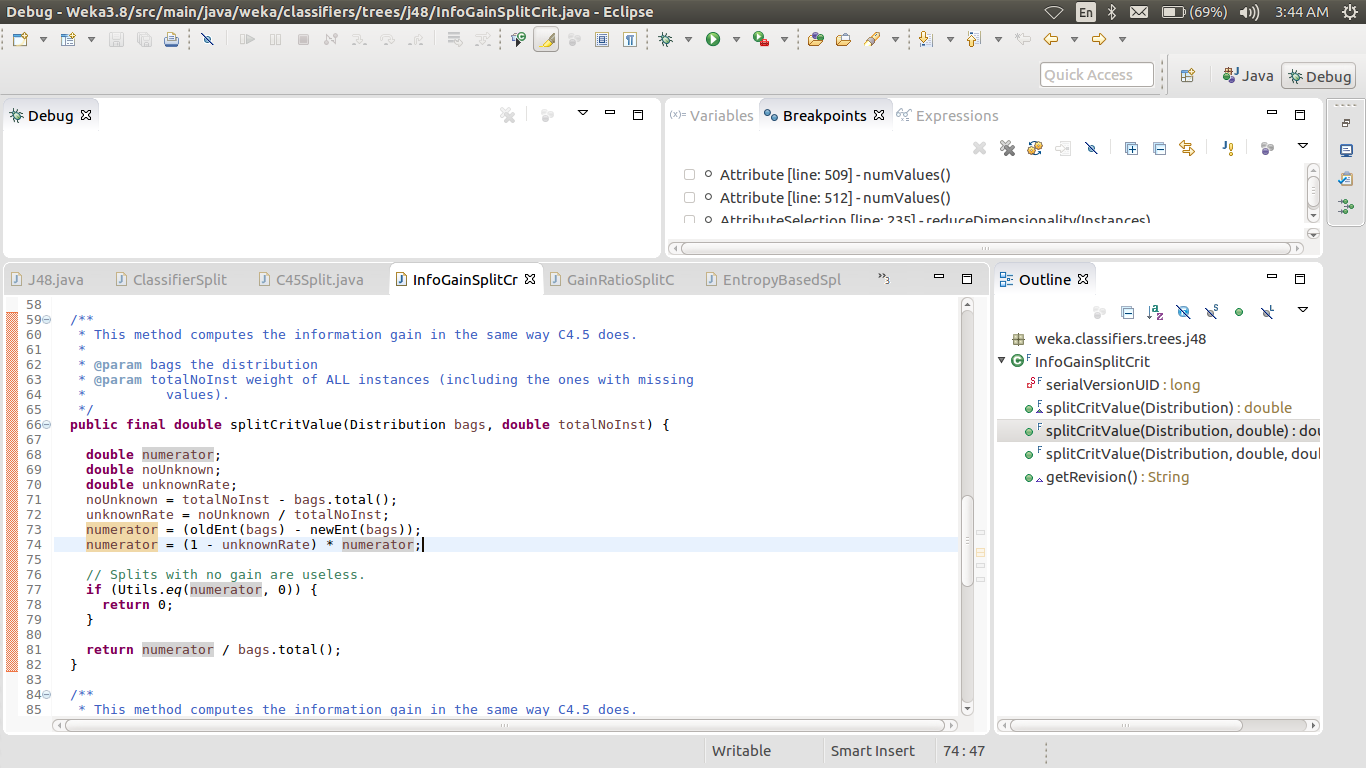
Firstly, the attribute name ‘PAY\_0’ was converted to ‘PAY\_1’. In the next step, all variables were numeric and hence some of them had to be converted to nominal. Also, values in the ‘EDUCATION’ variable was grouped such that the values ‘0,4,5,6’ was combined to one value and assigned a value ‘4’. For classification task, Target attribute are converted to nominal else it gives error.

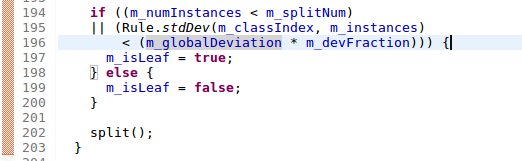
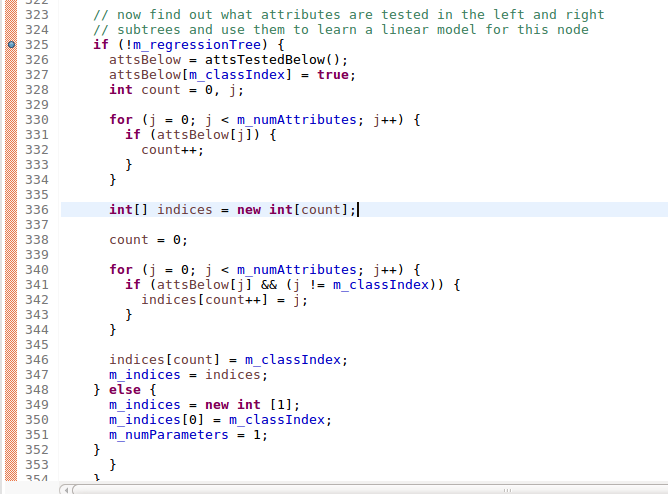
**Weka Code Description: Inputs, output, and process followed by Weka’s code to construct the trees (at most 2/3 page)**

**[10 points] J4.8 Code Description:**

First BuildTree Function is implemented. Over here, it checks whether instances belong to one class and if there are enough instances to split. It also, checks the basic parameters like whether the target is nominal or not. BuildClassifier function creates model for each attribute assuming that there are no missing values. For numeric data, it finds the spit point (variable= m\_splitpoint). Calculate Entropy for each split point using splitEnt function. It calculates the currentInfogain and finding gain for all splits, it selects the best gain. Info Gain= parent entropy – child entropy. Using best split point, split into subsets using Split function. This is the recursively repeated. For Nominal data, it calculates the entropy for each class in newEnt function. Based on that, InfoGain is calculated in splitCritValue function.

Best gain value is selected and that attribute is selected as splitting attribute. It checks whether the subsets have same classes and if not, recursively repeat. Else it’s a leaf node.

****

**[10 points] M5P Code Description: --**The algorithm receives all the instances of the data as the input in the beginning and calculates the standard deviation of the class. Splitting criteria is used to determine the best attribute to split the data. Based on the splitting criteria we create a node, and it then further divides the instances based on the split point. This process is called recursively until no further splits are possible. **--**WEKA makes use of “impurity” (expected error reduction) for partitioning the instances using the attribute. The instances with maximum impurity value is selected as the best split point. The splitting process terminates when the target values of the instances for the node has a standard deviation only slightly different than the standard deviation for the target attribute of the entire dataset. Also, splitting terminates when the number of instances is less than the threshold value. --WEKA uses the threshold value=4 i.e. when the number of instances are less than 4, the splitting terminates. An unpruned tree is received as output after this step. --The model is calculated for each node of the unprunned tree. Only the attributes tested below the particular node (below left and right subtree) are considered to build the model for the node.If “m\_useunprunned” flag in WEKA is set to false, the algorithm then starts pruning the tree from the leaf nodes as long as the expected estimated error for the linear model decreases. --The model considers only attributes from the subtree below the node and creates a model using these attributes, then it greedily drops each attribute as long as it improves the error estimate.

**[20 points] Python Packages and Functions used (decision trees, linear regression, model/regression trees). Describe inputs & outputs (at most 1/3 page)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Packages Used** | **Functions Used** | **Input Given** | **Output of code** |
| Decision Trees | Sklearn (model selection, metrics, tree), Time, pandas | Kfolds, predict, score, confusion\_matrix, roc\_auc\_score, DecisionTreeClassifier, predict\_proba, export\_graphviz, capacity, read | Credit data.csv | Tree which is printed using exportgraphviz. Error and Accuracy parameters |
| Linear Regression | Same as above, math, sklearn (linear\_model), matplotlib.pyplot | mean\_squared\_error, sqrt, LinearRegression, predict, score, fit, show, capacity, intercept\_, read | Credit data.csv | Model. Accuracy and error parameters. |
| Regression Trees | Same as above | mean\_squared\_error, sqrt, DecisionTreeRegressor, predict, score, fit, show, capacity, intercept\_ | Credit data.csv | Tree which is printed using export\_graphviz. Error and Accuracy parameters |

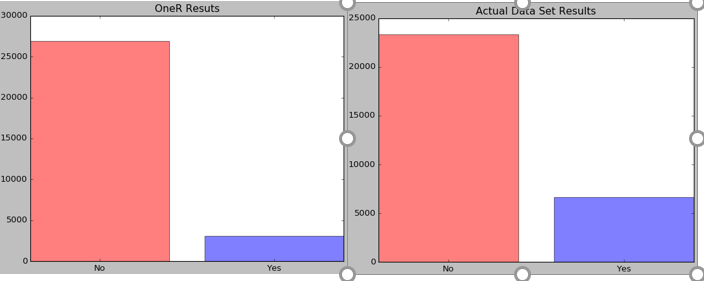
**[10 points] Three Guiding Questions for the Classification Experiments: (at most 1/3 page)**

1. **Influence of people who pay on time vs People who delays the payment in the month of September on defaulters- Hence, based on X6, one can predict whether his credit limit should be upgraded or degraded**
2. **Influence of Range of ages (age is for interpreting the results e.g. classify the results based on age groups) and history of past payment(X6-X11) on defaulters- Using age and past payment, we can find out the risk of being defaulter. Also, explained impact of CFS on accuracy and size of model**
3. **There is no Influence of Previous payment (X18-X23) on Defaulters- Even if the previous payment is bad or good enough, we cannot judge the person’s credit history just on the base of these parameters.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Classification Experiments in WEKA. Use 10-fold cross-validation** *At most 2/3 page.* | | | | | | | | | |
| **Tech.** | **Guiding ?** | **Pre-process** | **Parameters** | **Post-process &**  **Pruning** | **Accuracy,** **Precision, Recall, ROC Area** | **time** | **Size of model** | **Interesting patterns in the model** | **Analysis & observations about experiment** |
| ZeroR | 1, 3 | Removed all attribute other than target | Y | It requires only one column i.e. Target attribute (Y) to predict the results | Accuracy=77.77% Precision=0.607  Recall=0.779  ROC=0.5 | 0 sec | Predicted value or class | No change in Results of ZeroR even after adding attributes. | It takes the majority class (No) as the predicted output. Hence, accuracy for yes is 0. |
| OneR | 1 | - Converted the classes of X6 (Repayment status in September) to 1 (Person paid on time i.e. -2, -1, 0) and 2 (Payment delayed i.e. 1, 2 …) –> used RenameNominalValues in WEKA  -Nominal to Binary to X6 Column in WEKA  X7-X11 (Require only September data for repayment) | Sex, Education Marriage  X1-X6  X12-X17  X18-X23  Target as Y | First used with Pay\_1, then with Pay\_2 and then with Pay\_5. | ->81.96% of data are correctly classified  ->Precision=0.803  ->Recall=0.820  ->ROC Area=0.644 | 0.12 sec | 2 attributes (x6 and Y) are used to classify defaulters with 10 fold cross validation | Initially, it takes X6(pay\_1 Repayment in September) as the attribute to predict defaulters. Hence, we can answer the guiding question. Uses Misclassification metric to predict. | Increases accuracy and it has a better ROC area then the base method ZeroR. This is because it considers some data to predict the output. Also, with series of iterations we can gather that previous payments are extremely important. Accuracy reduced from 81.92 to 78.58% |
| J4.8 | 1,2 | Considered all attributes to compare the results with OneR. With all attribute, it is expected to give best accuracy | All | Pruning=True. | 80.11% correctly classified  Precision=0.78  Recall=0.80  ROC Area=0.64 | 4.61 sec | 1108 leaves and 2051 tree size | Even after increase in attributes, time and size its error parameters are more than One R | X6 is the more accurate predictor about defaulter compared to whole attributes. Also, X6 is the root node |
| J4.8 | 2 | -Discretize the age attribute in the bin of 3  -Removed ID (Redundant)  - Performing CFS results in age (Only when it is discretized) and X6-X11.  -Converted Age, Sex, Marriage Education X6-X11 to nominal (Done this for all experiments) | X6-X11  Age  Target(Y) | Pruned=true. | -> 82.08% of data are correctly predicted  ->Precision=0.804  ->Recall=0.821  ->ROC Area=0.691 | 0.78 sec | No. of leaves= 98  Tree Size= 187 | On the basis of age and history of past payment, model can accurately predict the target i.e. defaulters. | With such less number of attributes in input, accuracy is high enough compared to results of whole data. Less attribute (CFS) helps in reduce the tree size by almost 10 times. X6-X11 can be used to find the risk of being defaulter |
| J4.8 | 3 | -Removed ID(redundant)  -Performed correlation and removed attributes X18-X23 and Y (This has the lowest correlation) | Everything except X-18-X23, ID | Pruned=true | 81.19% accuracy  Precision=0.793  Recall=0.812  ROC Area=0.691 | 2.88 sec | No. of leaves= 569  Size= 1047 | Without past payment attributes, accuracy slightly increases. | Removing X18-X23 does not have much effect on target |
| J4.8 | 3 | -Removed ID(redundant)  -Performed correlation and removed all attributes except X18-X23 and Y (This has the lowest correlation) | X-18-X23  Y | Pruned=true | ->77.77% of data are correctly classified  ->Precision=0.607  ->Recall=0.779  ->ROC Area=0.5 | 0.23 sec | No. of leaves= 1  Tree Size= 1 | Results are similar to ZeroR. Also, ROC value is not worthy. | But when unpruned= true, Size of tree changes along with the ROC value which goes to 0.66. But results which are correctly classified is approx. same. Also, predictors have lowest correlation with the target |
| **[40 points] Summary of Classification Experiments in Python. Use 10-fold cross-validation** *At most 2/3 page.* | | | | | | | | | |
| ZeroR | 1,3 | -Extracted only the target | Target Column (Y) | Same as weka zeroR | Same as weka zeroR | 0.0019 sec |  | Same as weka zeroR | Same as weka zeroR |
| OneR | 1 | -using for loop, converted the classes of x6 to 1(Person paid on time i.e. -2, -1, 0) and 2 (Not paid on time) | Everything | Max depth=1  Criterion=Entropy | ->Score=0.8196  ->Precision=0.803  ->Recall=0.8196  ->ROC Area=0.6397 | 1.088 sec  Classifying X6 = 4.02 | Tree.node\_count=3L  One feature selected | Probability is high enough for predicted class. Accuracy (score) are approximately same for all folds | Attribute X6 is enough to predict accurately about defaulters. Probability of predicted class is high enough for each testing instance (predict\_proba). |
| J4.8 | 1 | -Removed ID | Used all except ID | Pruning= False (Because Accuracy is in question) | Score=0.733  Precision=0.730  Recall=0.612  ROC Area=0.643 | 20.203 sec | 7871 Nodes including leaf nodes | Even after taking all attributes, its accuracy is less as compared to OneR | Predicted class are either 1 or 0. In, some case its 0.5 for both classes. |
| J4.8 | 2 | -Discretize the age attribute in the bin of 3 using Numpy.histogram and for loop logic  -Removed ID  - Used Correlation to select highly correlated attribute to target(X6-X11) | X6-X11  Target(Y)  Age |  | -> Score=0.8156  ->Precision=0.797  ->Recall=0.649  ->ROC Area=0.678 | 60.80 secs (Discretizing age takes time) | 1729 Nodes including leaf nodes | Models accurately predicts the target with select attributes. Age is considered to refine the problem to age category. | Decrease in the number of attributes increases the accuracy. Hence, history of past payment is prominent factor to calculate the risk |
| J4.8 | 3 | -Removed ID(redundant)  -Performed correlation and removed attributes X18-X23(This has lowest correlation) | Everything except X-18-X23, ID | Pruned=False | -> Score=0.734  ->Precision=0.73  ->Recall=0.728  ->ROC Area=0.642 | 8.911 sec | 8693 Nodes including leaf nodes | Removing X18-X23 has no effect on accuracy and prediction | These attributes does not have much effect on target value |
| J4.8 | 3 | -Removed ID(redundant)  -Removed all attributes except X18-X23 and Y | X18-X23  Y | Pruned=False | -> Score=0.686 (Less than ZeroR)  ->Precision=0.68  ->Recall=0.53  ->ROC Area=0.5 | 6.004 sec | 1067 Nodes including leaf nodes | Results are worse than ZeroR with bad accuracy values | These attributes do not have much effect on target value. ROC value is worthless |

**[20 points] Summary of Weka and Python Classification Results, Analysis, Discussion, and Visualizations (at most 1/3 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best classification model obtained.

* Data Set has large number of No class in the target due to which even in worst condition, Number of No class are correctly predicted.
* X6 Attribute has very high correlation with the target. Even, OneR Method is taking X6 attribute as first question to predict the target. Hence, X6 alone is enough to predict the target accurately compared to all the attributes (Correlation Value= 0.32 which is high compared to other attributes). Moreover, just analyzing X6 attribute, Bank can answer various question like whether the person credit limit should be upgraded or whether he is eligible for loan.
* If Pay\_1 is not there in input, then it uses Pay\_2 or Pay\_5 as OneR attribute. Hence, bank can find many answers by just analyzing the history of past payment and can make necessary decisions. E.g. Whether clients credit limit or credit card should be upgraded or not.
* If we discretize age along with history of past payment, we can find various results about age group.
* X18-X23 has no effect because when only those attributes are considered in the model, then its results are worse than ZeroR
* OneR is the best classification model with less size and highest accuracy.
* All this functions are performed in python using KFolds (10-folds cross validation), mean square error (sklearn metrics), DecisionTreeClassification, export\_graphviz(To visualize the tree) and score, predict, tree\_.capacity, roc to analyze accuracy.
* If numeric to Nominal is not used on target, then it gives error
* If Numeric to Nominal is not done on X3-X11 then it creates less number of leaves as it considers them as continuous with very little decrease in accuracy

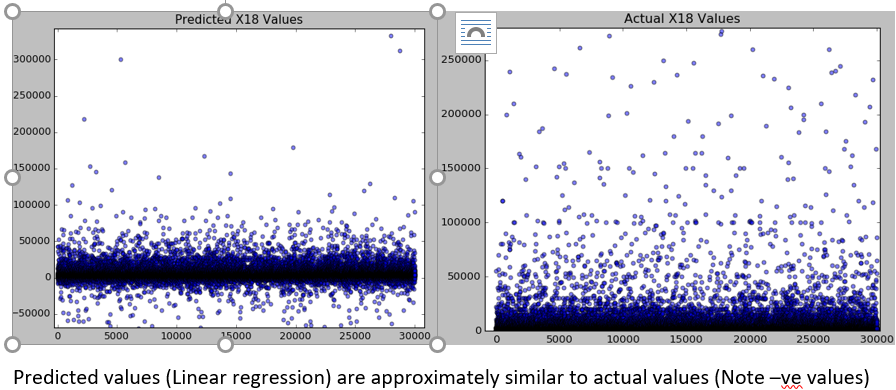


**[10 points] Three Guiding Questions for the Regression Experiments: (at most 1/3 page)**

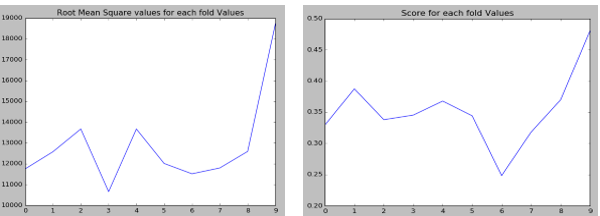
1. **According to the client’s history of payment and the bill amount for the month of August, how well can we predict if the client is going to pay the bill for September?**
2. **How much does the payment amount for each month affect the payment of a person in the month of September?**
3. **Does the social profile (eg: Sex, Marriage and Education) and evaluation of a client matter in deciding whether that particular client will pay in September or not?**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Classification Experiments in WEKA. Use 10-fold cross-validation** *At most 2/3 page.* | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Pre-process** | **Parameters** | **Post-**  **process**  **& Pruning** | **Correlation**  **Coefficient**  **and Error Metric(s)** | **Time to build model** | **Size of model** | **Interesting patterns in the model** | **Salient observations about experiment** |
| Linear  Regression | 1 | Added variables like Total\_Amount\_Paid (Sum of payment of months from X19-X23) | Total\_Amount\_Paid,  Bill\_Amt2 |  | Correlation coefficient 0.6178  Root mean squared error= 13024.3877 | 0.05 |  | The Bill Amount in September is used along with the total amount paid to predict how much the client will pay in the month of September | Root mean squared error reduced from 15110 to 13024. |
| M5P | 2 | -Removed Pay\_Amt2, Pay\_Amt3, Pay\_Amt4, Pay\_Amt5,  Pay\_Amt6 | All except Pay\_Amt2, Pay\_Amt3, Pay\_Amt4, Pay\_Amt5,  Pay\_Amt6 |  | Correlation coefficient 0.8241  RMSE 9383.13 | 9.37 seconds |  | Gives a high correlation coefficient when not including Pay\_Amt variables compared to ouput with whole data set (Correlation coef=0.422). | Interestingly one would wonder that the amount paid in the previous months by the client affects the payment for the September month but it did not. |
| M5P | 3 | -Removed Age, Sex, Education,Marriage  - Yes and No (Done for all experiments) | All except Sex, Education, Marriage, Age |  | Correlation coefficient: 0.9337  RMSE: 5937.8042 | 8.14 seconds |  | The correlation coefficient is close to 1 (0.93). Thus indicating that the social profile of the people does not help us in predicting the amount that a person will pay in the month of September 2005 | Removing these attributes from the model does not affect our decision to predict the amount paid in the month of September. |
| **[40 points] Summary of Classification Experiments in Python. Use 10-fold cross-validation** *At most 2/3 page.* | | | | | | | | | |
| Linear Regr | 1 | -Remove ID  -Change the class of Defaulters to 1 and 0 from | Everything except ID |  | R­­­­­2=0.352  RMS error=12896.30 | 0.45 sec | 25 attributes |  |  |  |  |
| Linear Regression | 1 | -Remove ID  -Change the class of defaulter to 1 and 0 from Yes and No (Pandas.replace)  -Change the education class to others(0,4,5,6) | Add a column named totalpaid= sum of payments.  Bill\_Amt2 and Pay\_amt1 | No Pruning | R2=0.342(Score Function)  Root Mean Square=12945.11 (RMS function) | 0.25 sec | 2 attributes in the model equation | Using linear regression model in scikit, We can predict the September payment just on the basis of 2 input columns | Ours model Accuracy is approximately same compared to the model which takes whole dataset as input. Hence, based on total payment and Bill of August, one can predict the payment in September. | Using linear regression model in scikit, We can predict the September payment just on the basis of 2 input columns | Ours model Accuracy is approximately same compared to the model which takes whole dataset as input. Hence, based on total payment and Bill of August, one can predict the payment in September. |
| Regression Tree | 2 | -Removed Pay\_Amt2, Pay\_Amt3, Pay\_Amt4, Pay\_Amt5,  Pay\_Amt6 | Everything except Pay\_Amt2, Pay\_Amt3, Pay\_Amt4, Pay\_Amt5,  Pay\_Amt6 | min\_samples\_leaf=3 | R­­­­­2=0.657  Root mean squared error=9489.19 | 29.81 | 12925 Nodes | Accuracy increases and error decreases compared to model with whole data as input | Interestingly one would wonder that the amount paid in the previous months by the client affects the payment for the September month but it did not. In fact, accuracy increases and error decreases | Accuracy increases and error decreases compared to model with whole data as input | Interestingly one would wonder that the amount paid in the previous months by the client affects the payment for the September month but it did not. In fact, accuracy increases and error decreases |
| RegrTree | 2,3 | Classified education as above | Everything except ID | min\_samples\_leaf=3 | R­­­­­2=0.50  RMS error=10940.53 | 42.45 sec | 12821 Nodes |  |  |  |  |
| Regression tree | 3 | -Removed Age, Sex, Education, Marriage | All except Sex, Education, Marriage, Age | min\_samples\_leaf=3 | R­­­­­2=0.57  Root mean squared error=10584.67 | 26 secs | 12919 Nodes | R2 value increases and RMS decreases which shows increase in accuracy | Removing these attributes does not affect our accuracy and hence these have no effect in predicting X18 | R2 value increases and RMS decreases which shows increase in accuracy | Removing these attributes does not affect our accuracy and hence these have no effect in predicting X18 |

**[20 points] Summary of Weka and Python Regression Results, Analysis, Discussion, and Visualizations (at most 1/3 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best regression model obtained.

The bill amount in the August along with the total bill amount paid over the past 5 months is used to find how much amount a person will pay in the month of September. The correlation coefficient obtained by using these attributes is reasonable enough to make this conclusion. Moreover, the social profile of the person has no significant impact on the person making a particular payment in the month of September. This can be seen from the model built without using Age, Sex, Marriage and Education. There is no significant difference in the RMSE and the correlation coefficient of the model vs the model built using the entire set of attributes.

Our best model is linear Regression model as it increases the correlation coefficient from 0.4 to 0.61 which is significant. Figure to the left shows the scatter plot of actual vs predicted target values (Linear regression). You can see that they are approximately similar but there are few values which are given negative amount as predicted value. That is because of too many 0 entry in actual values which we can see in graph.



Figures on left shows trend for RMS and Score for each fold and you can see that there are few RMS and score value which are very high. This shows that how cross validation helps in reducing the error due to selection of test and training sets. Python functions such as DecisionTreeRegressor, mean\_squared\_error (sklearn.metrics), KFolds (sklearn.model\_selection), LinearRegression and score , preditct, tree\_, capacity, export\_graphviz(To visualize the tree) were used to build the model

**Advanced Topic: Random Forest**

**[7 points] List of sources/books/papers used for this topic (include URLs if available):**

* http://jmlr.org/proceedings/papers/v28/denil13-supp.pdf
* Random Forests-Leo Breiman https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
* An Introduction to Statistical Learning with applications in R - Trevor Hastie, Robert Tibshirani, Daniela Witten, Gareth James

**[ 20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.**

**Random Forests**

Calculate the margin function based on the group of classifiers and the training set chosen at random from the distribution of the random vector.

Margin function measures the extent to which average number of votes for a predicted observation for the right class exceeds the average vote for any other class. The higher the margin the more confidence in the classification.

Random forest trees are constructed on the concept of bagging.

Random forest generates “n” trees for each “n” samples of the training data selected using bootstrap.

At each node, “x” predictor variables are selected at random.

Random forest uses only a subset of features to be considered while evaluating a split.

It then splits the data on the basis of the best feature from the subset.

At the next node, another “x” predictor variables are selected at random from all predictor variables and the above process is repeated until the leaf node is created.

The test-error of this model is calculated using OOB (Out of bag) error estimate.

We take average of each n predicted observations (if regression is the problem) or take the majority vote(if classification is the problem)

Random forest thus reduces variability of the prediction. Also, by increasing the number of bootstrap samples would not over-fit the data.

**[3 points] How does this topic relate to trees and the material covered in this course?**

Random Forest is an ensemble learning algorithm. The basic concept is that we build small number of weak decision trees with a subset of features selected at each node and then combining these trees to form a single strong learner by either averaging the predicted values of ith observation or taking the majority vote.

**Authorship:** Deepan Sanghavi (Data Observation, Python, WEKA Classification and Regression trees, Algorithm, Reporting), Dhaval Dholakia (Advance topic, Algorithm, WEKA regression, Python), Karan Napande (Data Exploration and Observation, WEKA Classification and Regression, PPT, Reporting)