CSE634 - Data Mining Concept and Techniques
Professor Anita Wasilewska

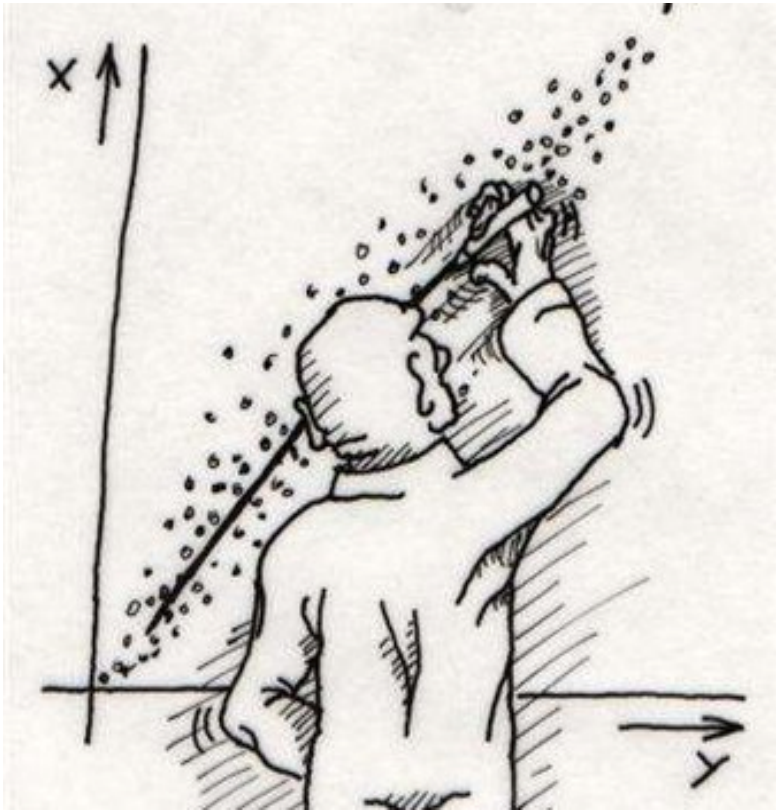
Regression Analysis

References

- https://pdfs.semanticscholar.org/7a07/5776db74495a03ca38750513f331b80f687e.pdf
- https://iterativepath.wordpress.com/2012/06/27/this-regression-model-is-beautiful-and-correctly-used/
- https://hbr.org/2015/11/a-refresher-on-regression-analysis
- https://algobeans.com/2016/01/31/regression-correlation-tutorial/
- http://dni-institute.in/blogs/scenarios-multiple-regression-applications/
- https://en.wikipedia.org/wiki/Polynomial_regression
- https://www.itl.nist.gov/div898/handbook/pmd/section8/pmd811.htm
- _Statistica- http://documentation.statsoft.com/
- http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis
- https://www.ndsu.edu/faculty/horsley/Stepwise_regression_(HZAU).pdf
- https://en.wikipedia.org/wiki/Mallows%27s_Cp
- https://people.duke.edu/~rnau/regstep.htm
- https://www.sciencedirect.com/science/article/pii/S0950705108001536
- https://en.wikipedia.org/wiki/Elastic_net_regularization
- http://web.stanford.edu/~hastie/TALKS/enet_talk.pdf

Content

Regression Analysis



Introduction

Image source : https://iterativepath.wordpress.com/2012/06/27/this-regression-model-is-beautiful-and-correctly-used/

What is Regression Analysis?

1. One of the most popular statistical method to understand the relation between variables of our interest better. For ex- Predicting a person's weight or how much snow we will get this year is a regression problem, where we forecast the future value of a numerical function in terms of previous values and other relevant features.
2. One continuous dependent (target) variable which we want to predict from number of independent (predictor) variables. How variation in one variable co-occurs variation in other.
3. We try to fit a curve/line to data points such that the differences between data point and curve is minimized.
4. Simple regression analysis in which one variable is used to predict another variable. Multiple regression analysis where multiple independent variables are used to predict the target variable.

Why we use Regression Analysis?

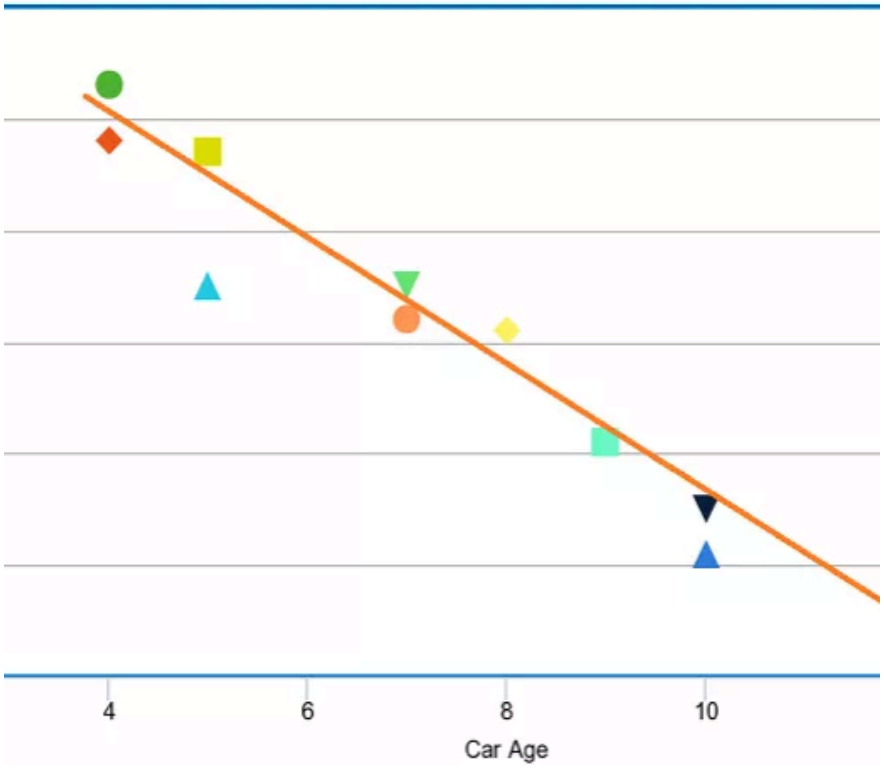
1. Quantify relationships and describe them using measured values and their graphical representation.
2. Provide forecasts and prediction

How Regression Analysis is done?

Let us consider an example where we want to predict the price of a used car sold by a car dealership company.

1. **Gather data on the variables:** Take all the used cars sold by company in last one year and data on any independent variable which we are interested in. Here we are considering the age of the car.
2. **Plot** all the information on the chart.
3. Now **draw a line that best fits the data** i.e. a line that runs roughly through the middle of the data (in scatter plot).

Image source: http://intellspot.com/scatter-plot/



In our example the variables are negatively correlated i.e. with increase in age of car the price of car decreases. We have only consider the car age for finding the price but in real world there can be number of variables which affect the price of car. For accurate results, we need to take into consideration all such variables.

Correlation and Regression?

When Data points are concentrated tightly along the line, it is sign that the predictor variable is strong and will be represented by large correlation coefficient. In the image, you can see as the magnitude of coefficient value is increasing, the scatter plot is getting concentrated along the line.

Correlation does not imply causation i.e. no cause and effect relationship. But in regression analysis builds cause and effect relationship.

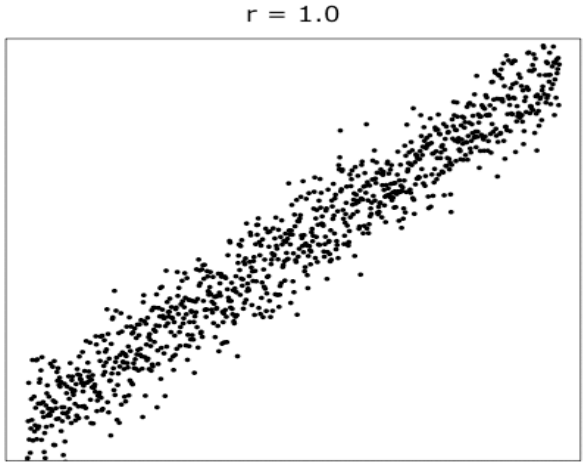


Image Source:https://algobeans.com/2016/01/31/regression-correlation-tutorial

The graph at top is showing different lines to fit in all the data points. The graph below that is the error in predictive accuracy.

We can see that, as the line is deviating from the best fit, the prediction error is increasing. Our main motive in Regression analysis is to find the best fit line for the data in order to minimize this error.

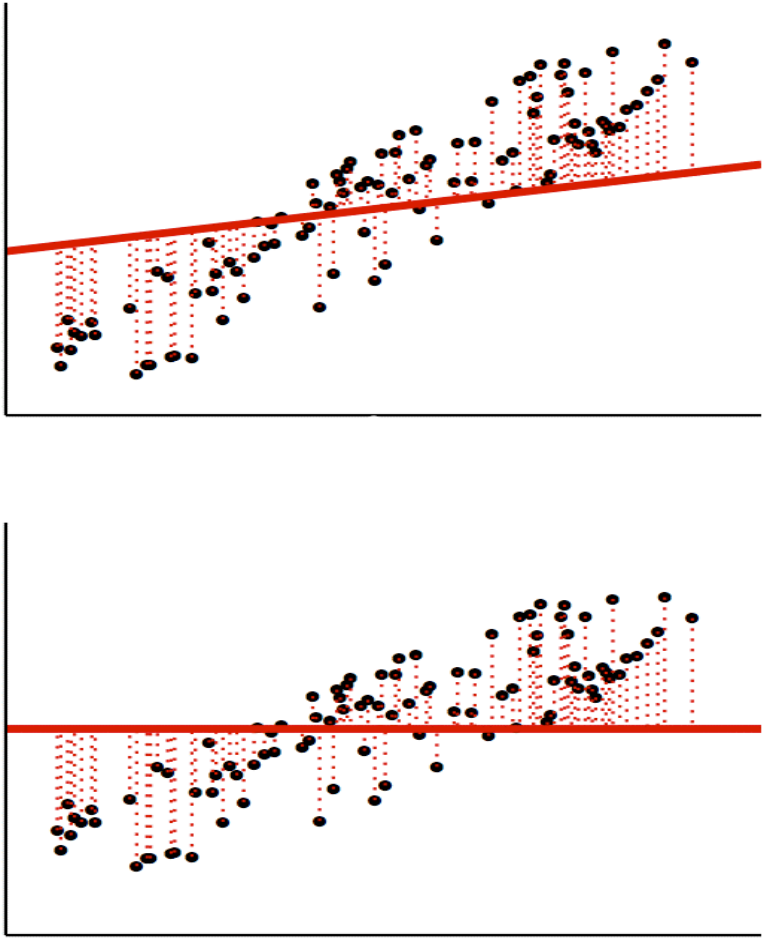


Image Source: https://algobeans.com/2016/01/31/regression-correlation-tutorial/

Application of Regression Analysis

1. Financial Forecasting
2. Crime Data mining: Predicting the crime rate of states based on drugs usage, human trafficking, killings etc.
3. Handwriting recognition
4. Software cost prediction, software quality assurance
5. Credit scoring
6. Measuring success rate of marketing strategy
7. Healthcare cost : Cost of healthcare for an individual to an insurer using claim history, demographic etc.
8. Salary estimate of a person

Types of Regression

Some of the widely used regression types are :

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. ElasticNet Regression

Polynomial Regression

Simple Linear Regression

$$Y = \beta_0 + \beta_1 x$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_n x_n$$

Polynomial Regression

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots\ldots + \beta_n x^n$$

* Where $\beta_1, \beta_2 .. \beta_n$ are unknown parameters

What is Polynomial regression?

- It is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as a nth degree polynomial in x. - wikipedia
- A general polynomial regression model is represented as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots\ldots + \beta_n x^n + \varepsilon$$

* ε is the random error factor.

Why do we need polynomial regression?

In simple linear regression, the model

$y = \beta_0 + \beta_1 x + \varepsilon$, is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable x. In this model, for each unit increase in the value of x, the conditional expectation of y increases by β_1 units.

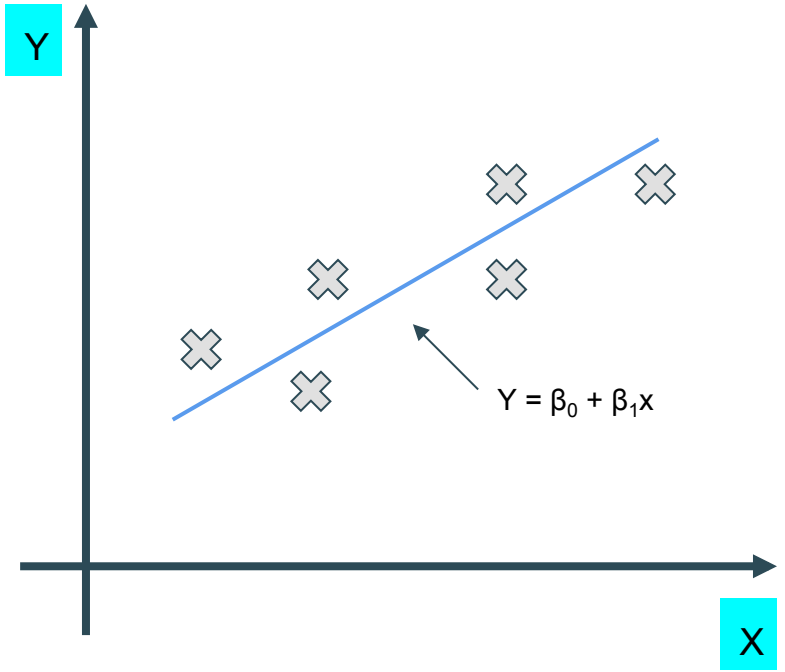


Image source: Self

Why do we need polynomial regression?

In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improved by increasing amounts for each unit increase in temperature.

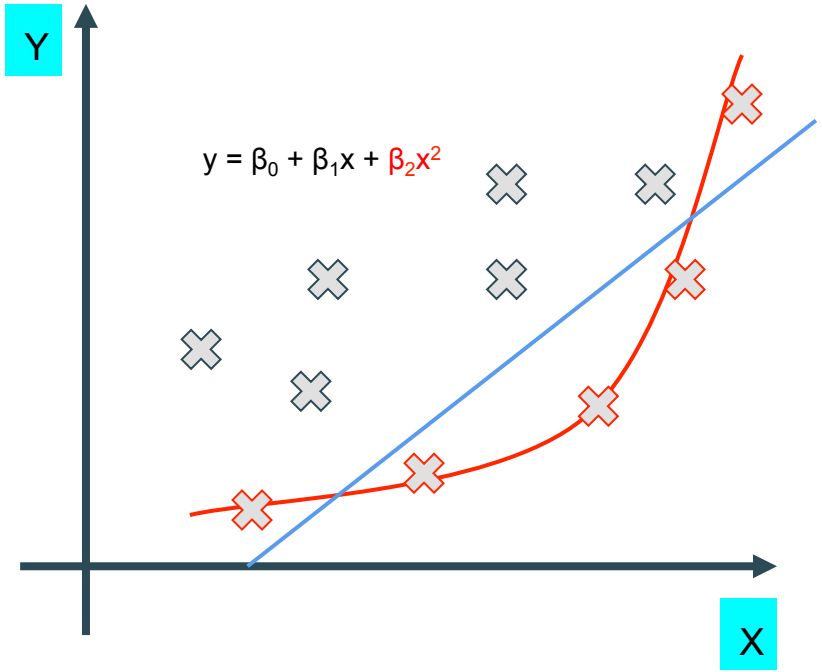


Image source: Self

Why do we need polynomial regression?

In such a case, we might propose a quadratic model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

- In this model, when the temperature is increased from x to x+1 units, the expected yield changes by $\beta_1 + \beta_2(2x+1)$.
- For a minute change in x, the effect on y is given by the total derivative with respect to x: $\beta_1 + 2\beta_2 x$.
- The fact that the change in yield depends on x is what makes the relationship between x and y nonlinear even though the model is linear in the parameters to be estimated.

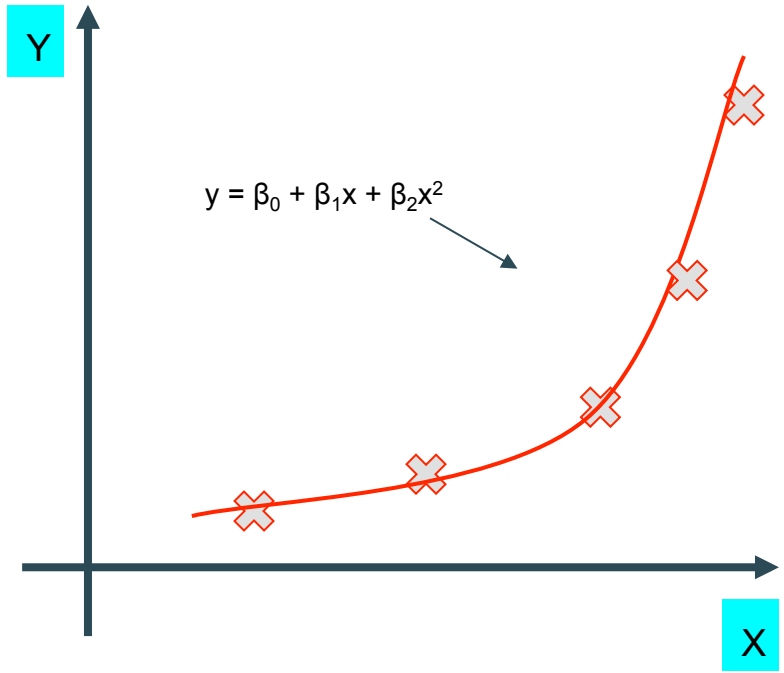


Image source: Self

How do we know that we need a polynomial regression to fit our model?

- Most important is the theoretical one. There are some relationships that researchers will hypothesize is curvilinear. Clearly, if this is a case, we include a polynomial term.

- The second chance is during visual inspection of our variables. We can plot a simple scatter plot that can reveal a curvilinear relationship indicating need of polynomial.

How do we know that we need a polynomial regression to fit our model?

- Inspection of residuals. If we try to fit a linear model to curved data, a scatterplot of residuals(y axis) on the predictor (x axis) will have patches of many negative residuals in middle, but patches of positive residuals at either end (or vice versa).

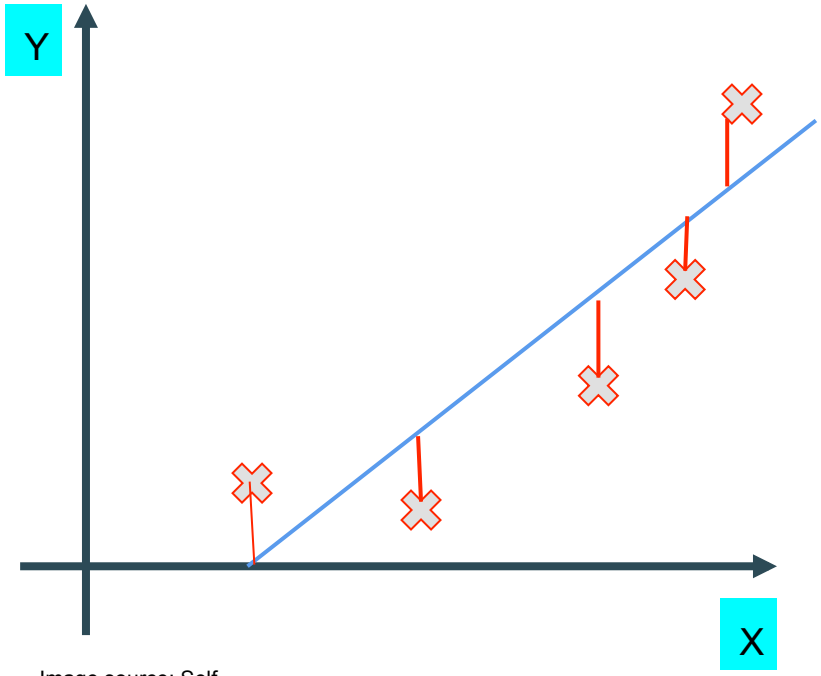
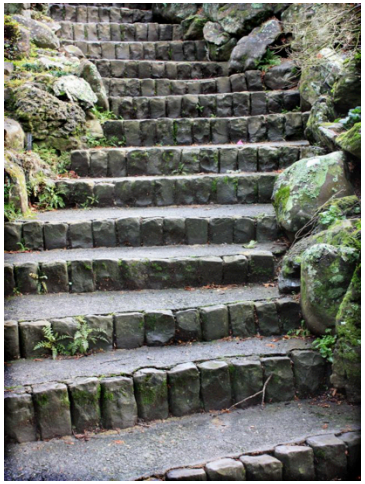


Image source: Self

Limitations to Polynomial Regression

- Polynomial models have poor interpolatory properties. High degree polynomials are notorious for oscillations between exact-fit values.
- They are sensitive to outliers. Polynomials may provide good fits within the range of data, but they will frequently deteriorate rapidly outside range of the data.
- Polynomial models have a shape/degree tradeoff. In order to model data with a complicated structure, the degree of the model must be high, indicating and the associated number of parameters to be estimated will also be high. This can result in highly unstable models.

Stepwise Regression



- A variable selection method for independent (predictor) variables where various combinations of variables are tested together.
- The “first step” will identify the “best” one-variable model. Subsequent steps will identify the “best” two-variable, three-variable, etc. models.
- Designed to find the most parsimonious set of predictors that are most effective in predicting the dependent variable.

Image source: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-smackdown-stepwise-versus-best-subsets

Selection criterion

- The “best” models are typically identified as those that maximize R2 , CP, or both.

R-square:

- **R-squared** is a statistical measure of how close the data are to the fitted regression line
- Goal in determining the best model is to minimize the residual mean square
- The model that contains all independent variables will give the maximum R2 value

Selection Criterion

CP criterion:

In statistics, **Mallows's C_p** named for Colin Lingwood Mallows, is used to assess the fit of a regression model that has been estimated using ordinary least squares. Small value of C_p means that the model is relatively precise.The CP values will decrease as the number of independent variables in the model increases.

- $C_P = (N - P - 1)\left(\frac{RMS}{\hat{\sigma}^2} - 1\right) + (P + 1)$, where
  - N = number of observations
  - P = number of independent variables in the model
  - RMS = residual mean square of the *P* selected variables
  - $\hat{\sigma}^2$ = is the residual mean square when all independent variables are included in the model.

Source: https://www.ndsu.edu/faculty/horsley/Stepwise_regression_(HZAU).pdf
https://en.wikipedia.org/wiki/Mallows%27s_Cp

How it works

- A combination of the forward and backward selection techniques.
- A modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. **If a nonsignificant variable is found, it is removed from the model.**
- Stepwise regression **requires two significance levels:** one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop.

How it works

Given is some set of potential independent variables from which we try to extract the best subset for use in your forecasting model.

At each step :-

For each variable currently in the model: compute the *t*-statistic for its estimated coefficient, *square* it, and report this as it's "*F*-to-remove" statistic;

For each variable *not* in the model: compute the *t*-statistic that its coefficient ***would*** have *if it were the next variable added*, square it, and report this as its "*F*-to-enter" statistic.

At the next step, the program automatically enters the variable with the highest *F*-to-enter statistic, or removes the variable with the lowest *F*-to-remove statistic, in accordance with certain control parameters specified by user.

Forward Selection

- A method of stepwise regression where one independent variable is added at a time that increases the R2 value.
- From group of variables that “can” be added, **add the one with the largest “variable added-last” t-statistic**.
- Addition of variables to the model stops when the “minimum F-to-enter” exceeds a specified probability level.

Backward Elimination

- A method of stepwise regression where all independent variables begin in the model and subsequent variables with least contribution are eliminated.
- Start with full model and delete variables that “can” be deleted, one by one, starting with the smallest “variable-added-last” t-statistic
- Elimination continues until the “minimum F-to-remove” drops below a specified probability level.

Forward Stepwise Regression

- Combine forward selection with backward elimination, checking for entry, then removal, until no more variables can be added or removed.
- Each procedure requires only that we set significance levels (or critical values) for entry and/or removal. Once this is done, each has exactly one result

Forward or Backward ?

- If we have a very *large* set of potential independent variables from which we wish to *extract* a few--i.e., if we're on a fishing expedition--we should generally go *forward*.

- If, on the other hand, we have a *modest-sized* set of potential variables from which we wish to *eliminate* a few--i.e., if we're fine-tuning some prior selection of variables, we should generally go *backward*.

Example

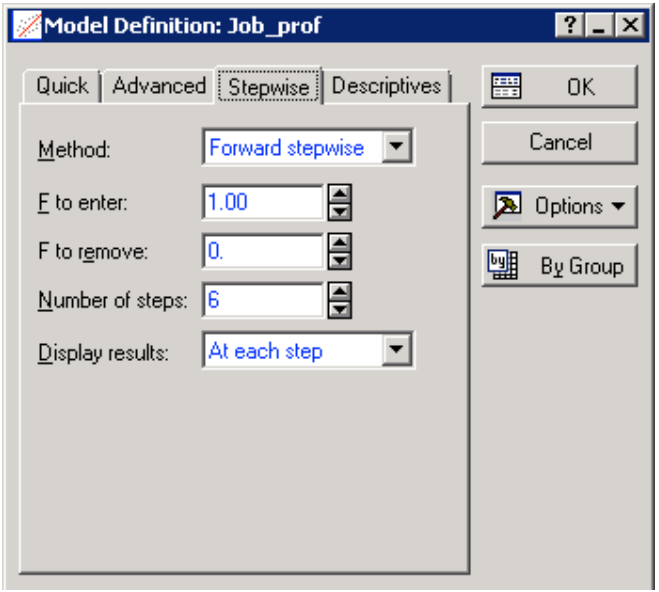
Analyzing using Forward stepwise regression :-

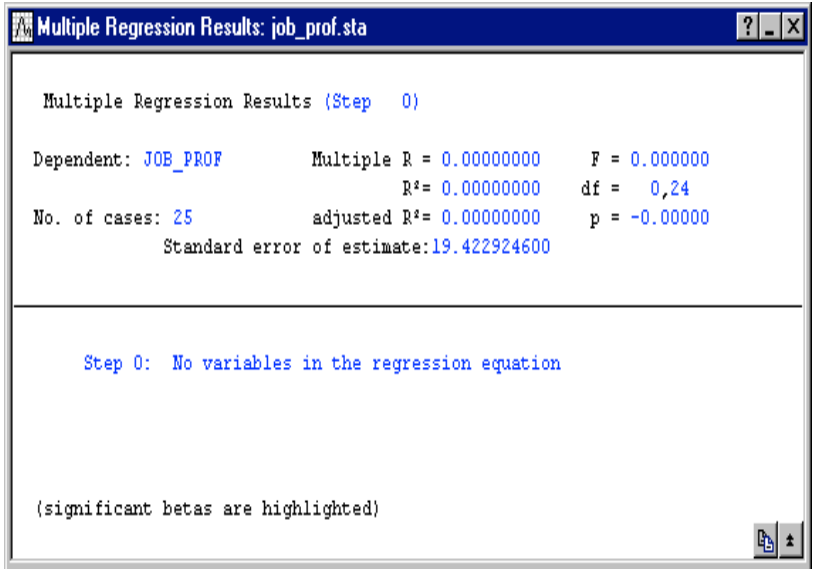
This example is based on the examples data file *Job_prof.sta* (from Neter, Wasserman, and Kutner, 1989, page 473). The first four variables (*Test1*-*Test4*) represent four different aptitude tests that were administered to each of the 25 applicants for entry-level clerical positions in a company. Regardless of their test scores, all 25 applicants were hired. Once their probationary period had expired, each of these employees was evaluated and given a job proficiency rating (variable *Job_prof*).

Research problem. Using stepwise regression, the variables (or subset of variables) that best predict job proficiency will be analyzed. Thus, the dependent variable will be *Job_prof* and variables *Test1*-*Test4* will be the independent or predictor variables.

Source:http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis

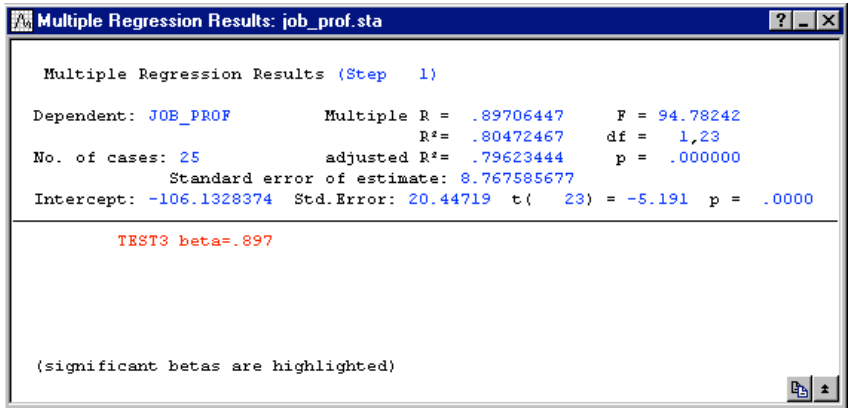
Example

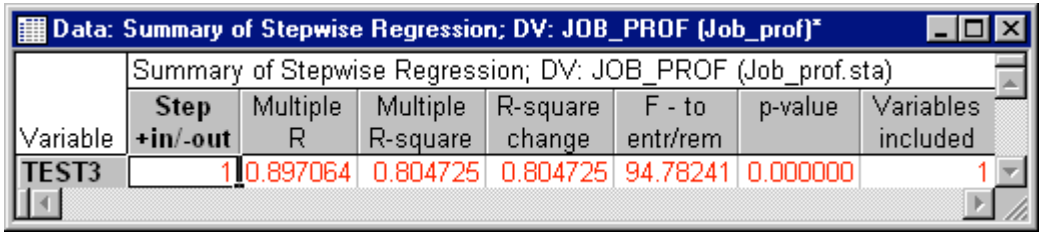




- Tool used: Statistica- http://documentation.statsoft.com/
- Source:http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis

Example





Variable	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-value	Variables included
TEST3	1	0.897064	0.804725	0.804725	94.78241	0.000000	1

- Tool used: Statistica- http://documentation.statsoft.com/
- Source:http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis

Example

Data: Summary of Stepwise Regression; DV: JOB_PROF (Job_prof)*

Summary of Stepwise Regression; DV: JOB_PROF (Job_prof.sta)

Variable	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-value	Variables included
TEST3	1	0.897064	0.804725	0.804725	94.78241	0.000000	1
TEST1	2	0.965917	0.932996	0.128271	42.11609	0.000002	2

Data: Summary of Stepwise Regression; DV: JOB_PROF (Job_prof)*

Summary of Stepwise Regression; DV: JOB_PROF (Job_prof.sta)

Variable	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-value	Variables included
TEST3	1	0.897064	0.804725	0.804725	94.78241	0.000000	1
TEST1	2	0.965917	0.932996	0.128271	42.11609	0.000002	2
TEST4	3	0.980583	0.961542	0.028547	15.58793	0.000735	3

- Tool used: Statistica- http://documentation.statsoft.com/
- Source:http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis

Example

The final regression equation is:

$y = -124.200 + 1.357*X_3 + 0.296*X_1 + 0.517*X_4$

Data: Regression Summary for Dependent Variable: JOB_PROF ...

Regression Summary for Dependent Variable: JOB_PROF (Job
R= .98058258 R²= .96154219 Adjusted R²= .95604822
F(3,21)=175.02 p<.00000 Std.Error of estimate: 4.0720

N=25	b*	Std.Err. of b*	b	Std.Err. of b	t(21)	p-value
Intercept			-124.200	9.874059	-12.5784	0.000000
TEST3	0.618670	0.069224	1.357	0.151832	8.9373	0.000000
TEST1	0.309670	0.045646	0.296	0.043679	6.7841	0.000001
TEST4	0.284405	0.072035	0.517	0.131054	3.9482	0.000735

- Tool used: Statistica- http://documentation.statsoft.com/
- Source:http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MultipleRegression/MultipleRegressionAnalysis/Examples/Example2StepwiseRegressionAnalysis

Applications

- Stepwise regression is an appropriate analysis **when you have many variables and you're interested in identifying a useful subset of the predictors.**
- Especially useful for sifting through large numbers of potential independent variables and/or fine-tuning a model by poking variables in or out.
- E.g:- predicting the college GPA using independent (predictor) variables, such as high school GPA and verbal SAT scores,
- Uses in economics, finance
- Predicting Life Expectancy Ratio by considering numerous variables like gender, eating habits, etc.

Advantages

- The ability to manage large amounts of potential predictor variables
- Fine-tuning the model to choose the best predictor variables from the available options.
- It’s faster than other automatic model-selection methods.
- Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables.

Limitations

- Stepwise regression often has many potential predictor variables but too little data to estimate coefficients meaningfully. Adding more data does not help much.
- If two predictor variables in the model are highly correlated, only one may make it into the model
- An inappropriate focus or reliance on a single best model

Limitations

- Inconsistencies among model selection algorithms
- Searches a large space of possible models. Prone to overfitting the data.
- Collinearity is usually a major issue. Excessive collinearity may cause the program to dump predictor variables into the model.
- Some variables (especially dummy variables) may be removed from the model, when they are deemed important to be included. These can be manually added back in.

Word of caution

Warning #1: *Be careful about including variables which have many fewer observations than the other variables :-*

Because the stepwise algorithm uses a correlation matrix calculated in advance from the list of all candidate variables, therefore, be careful about including variables which have many fewer observations such as seasonal lags or differences

As they will shorten the test period for all models whether they appear in them or not, and regardless of whether "forward" or "backward" mode is used.

Word of caution

Warning #2: *Excessive multicollinearity:*-

If the number of variables that selected for testing is *large* compared to the number of observations in the data set (say, more than 1 variable for every 10 observations),

or if there is excessive multicollinearity (linear dependence) among the variables

then the algorithm may go crazy and end up throwing nearly all the variables into the model.

Word of caution

Warning #3: *Critical variables thrown out:*-

Sometimes we may have a subset of variables that ought to be treated as a group (say, dummy variables for seasons of the year) or which ought to be included for logical reasons.

Stepwise regression may blindly throw some of them out, in which case we should manually put them back in later.

Word of caution

Warning #4: *Do your homework before you begin:-*

Automated regression model selection methods only look for the most informative variables from among those we start with, and *they cannot make something out of nothing*.

For example- if the assumption of linear or linearizable relationships is simply wrong, no amount of searching or ranking will compensate.

The most important steps in statistical analysis are (a) doing your homework before you begin, and (b) collecting and organizing the relevant data.

Elastic Net Regression: Why?

- When we are working with high dimensional data (datasets with a large number of independent variables), correlations between the variables can be high resulting in multicollinearity.
- These correlated variables can sometimes forms groups or clusters of correlated variables.
- There are many times where we would want to include the entire group in the model selection if one variable has been selected.
- This can be thought of as an elastic net catching a school of fish instead of singling out a single fish.

Elastic Net Regression: What?

- Penalized regression

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

- Hybrid of Ridge and Lasso Regression

Types of Penalized Regressions

- Ridge regression
- LASSO
- Elastic net

Ridge Regression

- Penalty is an additive term proportional to the sum of squares of coefficients
- Does not $\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \}$ used for variable selection and is not easy to interpret

Lasso Regression

- Lasso is an acronym for “Least Absolute Selection and Shrinkage Operator”.
- Data (X, y). X is the n × p predictor matrix of standardized variables; and y is the response vector.

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \quad \text{s.t. } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t$$

- Cons: If a group of pr ick only one of them and shrink the other to zero , can not do grouped selection.

Elastic Net Regression

- Elastic net is a regression method used to simplify Linear and Logistic regression models through regularization
- Linearly combines the L_1 and L_2 penalties of the Lasso and Ridge methods.
- It is trained with L1 and L2 prior as regularizer.
- Elastic-net is useful when there are multiple features which are correlated.

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

Elastic Net Regression (Cont)

- The total number of variables that the lasso variable selection procedure is bound by the total number of samples in the dataset.
- Additionally, the lasso fails to perform grouped selection. It tends to select one variable from a group and ignore the others.
- The elastic net forms a hybrid of the ℓ1 and ℓ2 penalties:

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1$$

$$\hat{\beta}^{\text{elastic}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

- The L_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty (L_2)
  - Removes the limitation on the number of selected variables;
  - Encourages grouping effect;
  - Stabilizes the L_1 regularization path.

In a nutshell...

- **Ridge Regression:**
  - good for multicollinearity, grouped selection
  - not good for variable selection
- **LASSO**
  - good for variable selection
  - not good for grouped selection for strongly correlated predictors
- **Elastic Net** - combine strength between Ridge Regression and LASSO

Research Paper

Feature selection in bankruptcy prediction

Department of Information Management, National Central University, 300 Jhongda Road, Jhongli 32001, Taiwan

Received 9 January 2008, Revised 14 July 2008, Accepted 7 August 2008, Available online 14 August 2008.

https://doi.org/10.1016/j.knosys.2008.08.002

Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

What is Bankruptcy?

Bankruptcies are companies/individuals which can not operate continually or get the awful credit.

Bankruptcy occurs if the company can not operate, pay liability, earn profits and obtain bad credits, etc.

Forecasting bankruptcy can be thought of as a classification problem. With input variables as the financial and accounting data of a firm, we try to find out which category the firm belongs to bankruptcy or non-bankruptcy.

Why it’s an important problem?

For many corporations, assessing the credit of investment targets and the possibility of bankruptcy is a vital issue before investment. Data mining and machine learning techniques have been applied to solve the bankruptcy prediction and credit scoring problems.

Related Work

In the past, Beaver uses financial ratios as the input variables for linear regression models to classify healthy/bankrupt firms.

Altman [11] uses the classical multivariate discriminate analysis technique. On the other hand, many recent studies focus on using data mining techniques [12] for bankruptcy prediction Related work shows that data mining models (e.g. neural networks) outperform statistical approaches (e.g. Logistic regression, linear discriminate analysis, and multiple discriminate analysis) [1,13–16].

Source:https://www.sciencedirect.com/science/article/pii/S0950705108001536

Why DM in Bankruptcy Prediction

To deeply analyze a huge amount of information of the corporations is likely to take much time and need many human resources. When irrelevant information is overabundance, it is unlikely to interpret and absorb the information very easily. Therefore, how to filter and condense the large amount of data is a very important issue to predict business failures, especially for bankruptcy prediction.

Feature selection as the preprocessing step is the one of the most important steps in data mining process. It aims at filtering out redundant and/or irrelevant features from the original data

Human Bias can also lead to biased results.

Feature Selection

Feature Selection is an important aspect which is often ignored. It is important to choose a group of set of attributions with more prediction information. Reducing the number of irrelevant or redundant features drastically reduces the running time of a learning algorithm and yields a more general concept.

It is unknown that which feature selection method is better.

Therefore, this paper aims at comparing five well-known feature selection methods used in bankruptcy prediction, which are *t*-test, correlation matrix, stepwise regression, principle component analysis (PCA) and factor analysis (FA) to examine their prediction performance. Multi-layer perceptron (MLP) neural networks are used as the prediction model. Five related datasets are used in order to provide a reliable conclusion.

Methods in Feature Selection

1. **Correlation matrix** - Correlation matrix is to confer the correlation of two quantitative groups, as well as to analyze whether one group affects the other one. The relationship between two variables is said to be highly correlated if a movement in one variable results or takes place at the same time as a similar movement in another variable

2. ***T*-test** - The *t*-test method is used to determine whether there is a significant difference between two group's means. It helps to answer the underlying question: do the two groups come from the same population, and only appear differently because of chance errors, or is there some significant difference between these two groups

Methods in Feature Selection (cont)

3. **Factor analysis** - The purpose of FA is to describe concisely the interrelationship of the numerous variables.FA seeks the least number of factors which can account for the common variance (correlation) of a set of variables

4. **Principle component analysis** - The central idea of PCA is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, as the principal components, which are uncorrelated and ordered so that the first few retain most of the variation present in the entire original variables.

5.**Stepwise regression**- The most common technique to find the best combination of predictor variables is stepwise regression. Although there are many variations, the most basic procedure is to find the single best predictor variable and add variables that meet some specified criterion.

Predictive Accuracy

After using 5-fold cross-validation, we can evaluate which model is the most appropriate model as the baseline, i.e. provides the highest prediction accuracy.

Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Experiment Design

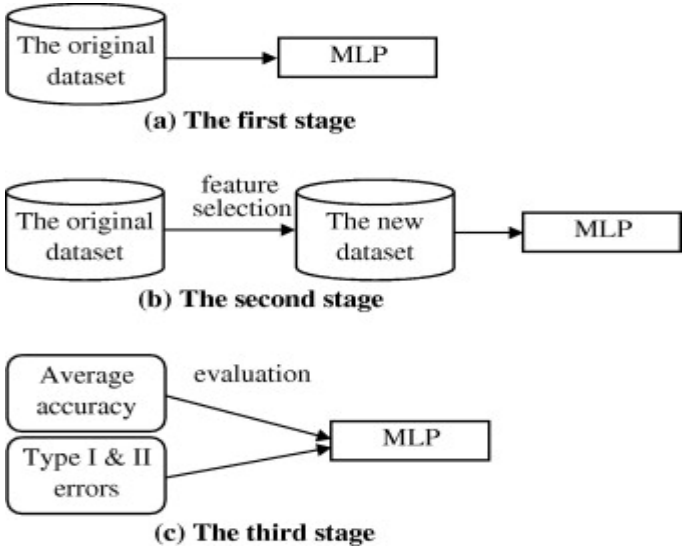


Image Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Types of Error

Type I error

It means that the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when one does not have adequate power. That is, in bankruptcy prediction it occurs when we classify the non-bankruptcy group into the bankruptcy group.

Type II error

It means that the error of rejecting a null hypothesis when it is the true state of nature. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when an observation is due to chance. In bankruptcy prediction, it occurs when we classify the bankruptcy group into the non-bankruptcy group.

	Japanese credit	Australian credit	Bankruptcy dataset	German credit	UC competition
Learning epoch	400	400	400	100	50
Hidden nodes	64	32	8	16	32
Average Accuracy	85.88%	81.93%	71.03%	74.28%	96.92%
Type I error	90.05%	21.89%	12.85%	55.39%	81.68%
Type II error	22.40%	13.89%	30.42%	9.63%	4.05%

Image Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Table 5. Performance of feature selection unit: %

	t-test	Stepwise	Correlation matrix	FA	PCA	Baseline models	*F*value
Japanese credit							
Accuracy	63.53	82.64	60.16	74.22	74.00	85.88	3.25*
Type I error	55.33	32.27	74.55	29.17	47.46	90.05	3.521*
Type II error	17.29	6.77	3.49	23.75	10.37	22.40	2.046
Australian credit							
Accuracy	89.27	84.74	89.31	86.08	89.93	81.93	7.279**
Type I error	9.38	12.80	13.33	14.58	7.93	21.89	7.949**
Type II error	11.72	16.71	8.33	13.60	11.53	13.89	7.136**

Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Feature selection Method Ranks

By evaluating means of the results, we can rank the five feature selection methods by their performance which has significant difference.

Therefore, for average accuracy, the first one is factor analysis and than *t*-test. For the Type I error, the first place is *t*-test and the second one is stepwise. Correlation matrix and stepwise are the first and second methods to effectively reduce the Type II error, respectively.

To sum up, *t*-test is the better feature selection method to provide higher prediction accuracy and reduce the Type I error. On the other hand, stepwise extracts the most features (see next subsection), and provides relatively better performance in terms of average accuracy.

Feature Selection Method Ranks

Table 7. Ranking results

	Number 1	Number 2
Average accuracy	Factor analysis	*t*-test
Type I error	*t*-test	Stepwise
Type II error	Correlation matrix	Stepwise

Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Feature Extraction Rate vs Accuracy

Fig. 2 shows the feature reduction percentage versus its prediction performance.

The results show that stepwise diminishes substantially the number of redundant or irrelevant features, but the accuracy rates are not the worst one over the four datasets.

On the other hand, FA reduces the lowest rate of irrelevant features. It provides the best result in the German data set, but has the worst result in the Bankruptcy dataset. For *t*-test, it lies on the middle position of the feature reduction rate and perform very good over the four datasets.

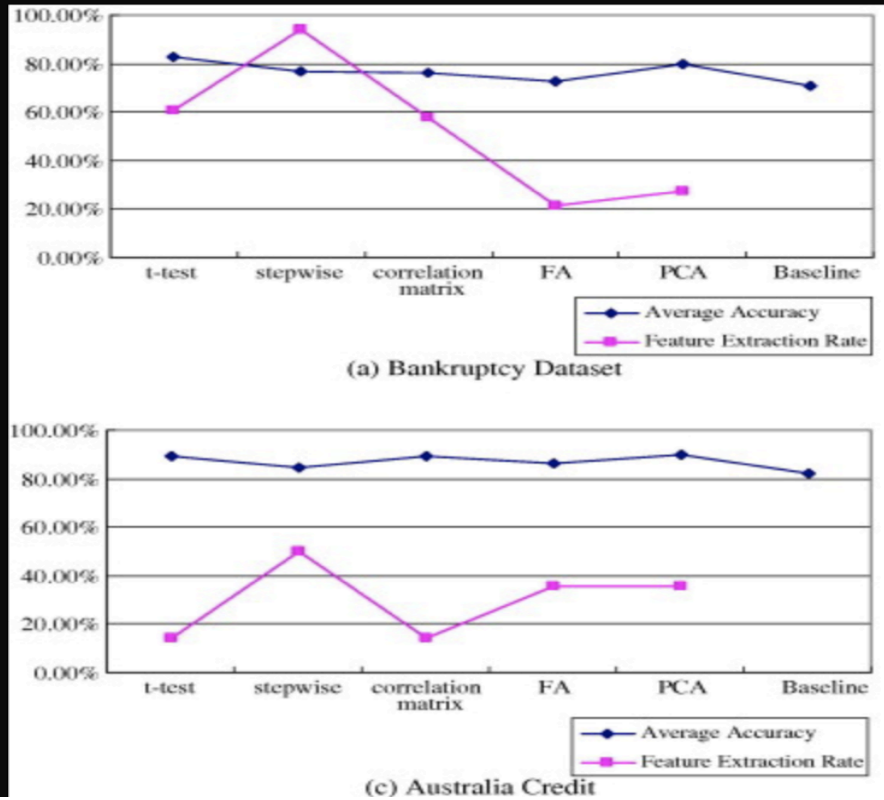


Image Source - https://www.sciencedirect.com/science/article/pii/S0950705108001536

Conclusion

Regarding the experimental results, feature selection methods applied on selecting more representative variables certainly increase the performance of prediction. On average, *t*-test is superior to others and Stepwise is on the second position. For the percentage of reducing the original variables, stepwise outperforms the others, which provides the highest feature reduction rate.In summary, *t*-test performs stably and provides higher prediction accuracy and lower Type I and II errors.

For future work, the research findings could be considered in other related business domains, especially for two-class classification problem which is the same as the bankruptcy prediction problem, such as stock price prediction (stock up and down), customer churn prediction (churn and non-churn), etc.