Institute of Technology
Blanchardstown
Institiúid Teicneolaíochta
Baile Bhlainséir

MSc in Computing, Business Intelligence and Data Mining stream.

Business Intelligence and Data Mining Applications Project Report.

Predicting earning potential on Adult Dataset

Submitted by: xxxxxxx

Supervisor: Markus Hofmann

Submission date 21/05/2011

Declaration

I herby certify that this material, which I now submit for assessment on the programme of study leading to the award of MSc in Computing in the Institute of Technology Blanchardstown, is entirely my own work except where otherwise stated, and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfillment of the requirements of that stated above.

Author: __xxxxxxxx______ Dated: __21/5/2011___

Abstract

This report implements the CRISP-DM methodology when applying classification models to the problem of identifying individuals whose salary exceeds a specified value based on demographic information such as age, level of education and current employment type. The process involved in the exploration, preparation, modelling and evaluation of the datasets are described. Topics such as the application of statistical analysis to suggest attribute usefulness, feature reduction, outlier detection, missing value management, data bias and data transformation are discussed. The process of relative performance analysis of the proposed classifiers is reviewed. The support of a business objective which will use the predictive capabilities of the proposed models to target customers is reviewed including the use of lift analysis to indicate the likely level of return on investment and overall profitability.

Table of Contents

	Page
Introduction	6
Business Understanding	6
Business objective	6
Data Mining objective	7
Project plan	7
Data Understanding	7
Describe the data	7
Explore the data	8
Verify data quality	14
Data Preparation	15
Select Data	15
Construct Data	16
Modelling	19
Select modelling technique	19
Generate Test Design	19
Build and Assess the model	19
Evaluation	28
Appendix A	30
References	30

List of Tables

1	Description of dataset	9
2	Numerical attribute quantiles, unique values and outliers	11
3	Mapping between education number and education level attributes	11
4	Comparison of occupation attribute values within the salary classes	13
5	Rules discovered for >50K salary class on training dataset	14
6	Naïve Bayes with forward selection, SVD and PCA	16
7	Relationship and marital status attribute transformations	17
8	Entropy binning output	17
9	Numerical attribute binning	18
10	Naïve Bayes performance (iteration 2)	20
11	Naïve Bayes and hybrid classifier performance	20
12	Discretisation applied to NBTree classifier	22
13	kNN k values	22

14 kNN performance 22
15 Mean values cost attribute clusters 25
16 Classifier evaluation on test dataset 25
17 Classifier Lift results 26

List of Diagrams

1 Boxplots of numeric attributes 10
2 Regression classifier output 11
3 Training dataset correlation matrix 12
4 Histogram of education levels 12
5 Histograms of age for gender and salary 12
6 Decision tree output 13
7 Outlier deviation plot 14
8 Initial ROC curves 15
9 Data transformations type A 18
10 Data transformations type B 19
11 Hybrid classifier ROC curves 20
12 NBTree classifier output 21
13 Data transformations type C 23
14 Hybrid ROC curves on transformed data 23
15 MetaCost scatter plot 24
16 Return on investment plots 27
17 Profitability plots 27

1. Introduction

This project will investigate the data mining of demographic data in order to create one or more classification models which are capable of accurately identifying individuals whose salary exceeds a specified value. The data used in this project were sourced from the University of California Irvine data repository and are referred to as the Adult dataset and contain information on individuals such as age, level of education and current employment type.

The classification model will be used to select candidates for a new service offered by the sponsor of the project targeting individuals with salaries exceeding fifty thousand US dollars. A description of the predictive significance of each attribute, interesting or useful patterns which were found and any transformations applied to the data must be provided with all proposed models.

This report will describe the work carried out during the iterative process of data preparation, modelling and evaluation including data formatting, consistency or other quality issues, opportunities for useless instance or attribute removal and the approaches taken to solving issues with instances affected by noise, outliers or missing values.

This project will implement the CRISP-DM methodology where a comprehensive review of the customer's requirements supports the creation of a business objective outlining items such as the expected level of model performance and return on investment. This will be used to create a data mining objective which will guide the subsequent work in the data understanding, preparation and modelling steps and the final evaluation and selection (after revisiting earlier steps if necessary) of a classification model or models.

2. Business Understanding

Business objective

A project team has been created to support the marketing of a new service targeted at potential customers with medium to high level salaries. There is an initial project setup cost of eighteen thousand dollars, a cost of one hundred and twenty five dollars for each offer made and a return of five hundred dollars for each accepted offer:

Setup cost:	$18,000
Cost per offer:	$125
Return per accepted offer:	$500

The current marketing strategy, which involves a high degree of investment per offer (driving the relatively high offer cost) has achieved a high acceptance rate in the past of seventy five percent by individuals whose salaries exceed fifty thousand US dollars. The goal of this project is therefore to create a model (or models) which can accurately identify individuals whose annual salary exceeds this amount. Any proposed model must be capable of significantly outperforming the existing candidate selection model which is currently providing an average return on investment of seventy to ninety.

Data Mining objective

The data mining objective is to create a classification model which can predict individuals whose salary exceeds fifty thousand US dollars by mining anonymised census data containing demographic information such as age, gender, education level and employment type. The original salary attribute in the census data has been anonymised to a binomial value indicating if a salary exceeds fifty thousand US dollars.

For each proposed model an expected return on investment (and associated profit margins) must be provided. A clear description of all data transformations which were carried out must be provided and any useful insights into the data such as significant attributes or mining issues within the data should be included.

Project plan

Phase	*Date*	*Details*	*Status*
A	20/2/2011	Start date	*Closed*
B	28/2/2011	Project proposal compete	*Closed*
C	5/3/2011	*Crisp 1*: Business understanding	*Closed*
D	12/3/2011	*Crisp 2*: Initial data understanding	*Closed*
E	20/3/2011	*Crisp 3*: Initial data preparation (and testing with early modelling)	*Closed*
F	30/03/2011	Interim presentation based on phases A to D	*Closed*
		Crisp 4: Creation/test of models (update results in D and E as required)	*Closed*
G	06/04/2011	*Crisp 4*: Ongoing testing of data preparation (update results in D to F as required)	*Closed*
H	20/04/2011	*Crisp 5*: Evaluation of proposed models	*Closed*
I	1/5/2011	Final presentation of work	*Closed*
		Create project report	*Closed*
J	16/5/2011	Completion date	

3. Data Understanding

Describe the data

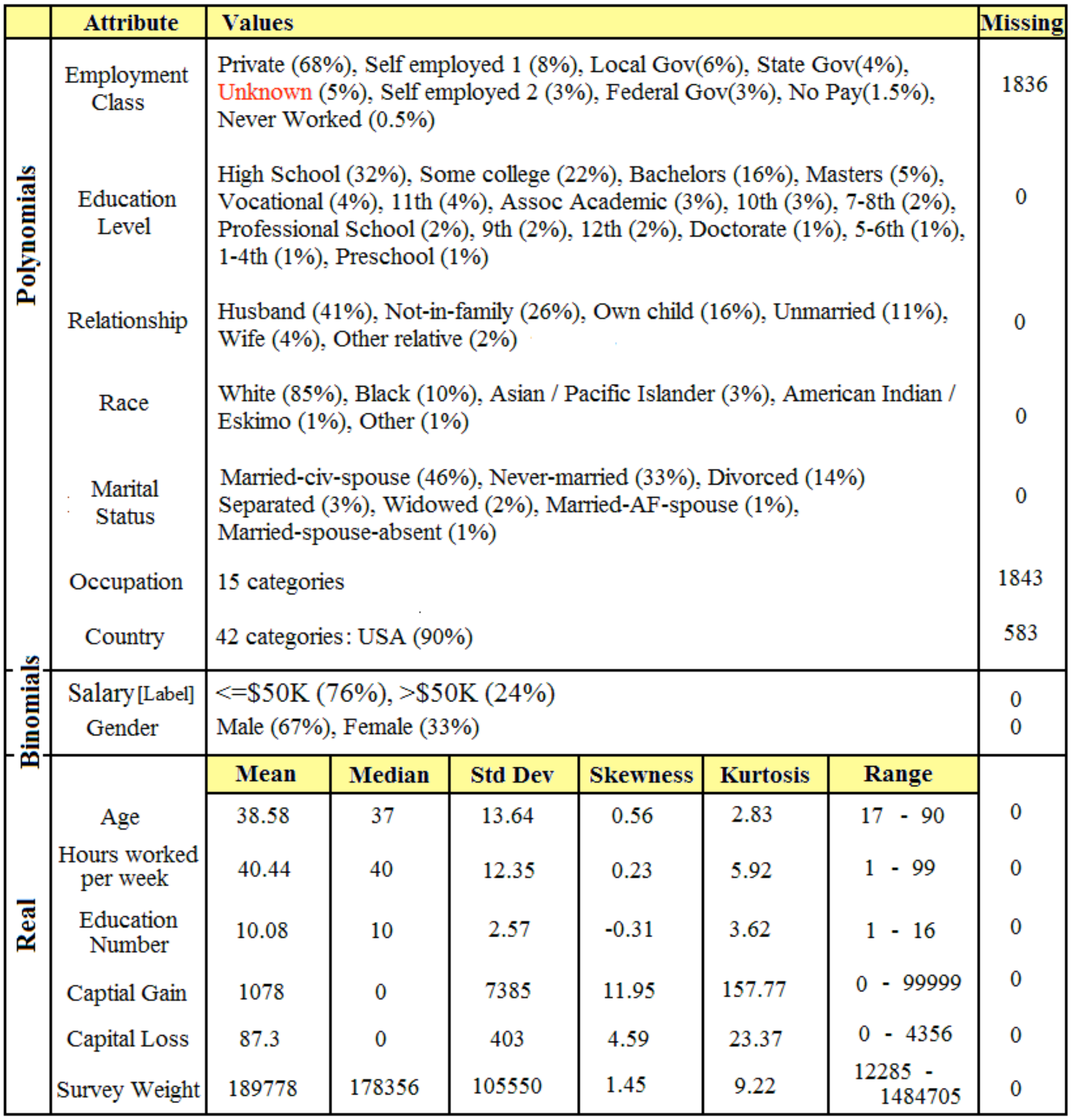
The dataset used in this project has forty nine thousand records and a binomial label indicating a salary of less or greater than fifty thousand US dollars, which for brevity, will be referred to as <50K or >50K in this report. Seventy six percent of the records in the dataset have a class label of <50K. The data has been divided into a training dataset containing thirty two thousand records and a test dataset containing sixteen thousand records.

There are fourteen attributes consisting of seven polynomials, one binomial and six continuous attributes (Table 1). The nominal employment class attribute describes the type of employer such as self employed or federal and occupation describes the employment type such as farming or managerial. The education attribute contains the highest level of education attained such as high school graduate or doctorate. The relationship attribute has categories such as unmarried or husband and the marital status attribute has categories such as married or separated. The final nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (which is a numerical representation of the nominal education attribute), capital gain and loss and a survey weight attribute which is a demographic score assigned to an individual based on information such as area of residence and type of employment.

Explore the data

The standard deviations in the data (Table 1) indicates that there is a significant quantity of values in all attributes, particularly in the case of the survey weight attribute and the high kurtosis in the capital gain attribute indicates a long tail. From the boxplots it can be seen that the range of attribute values for age, education number and hrs_per_week (worked) is slightly higher in >50K class instances (Figure 1) indicating that these attributes may have predictive significance.

Table 1. Description of dataset.



	Attribute	Values						Missing
Polynomials	Employment Class	Private (68%), Self employed 1 (8%), Local Gov(6%), State Gov(4%), Unknown (5%), Self employed 2 (3%), Federal Gov(3%), No Pay(1.5%), Never Worked (0.5%)						1836
	Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%), Vocational (4%), 11th (4%), Assoc Academic (3%), 10th (3%), 7-8th (2%), Professional School (2%), 9th (2%), 12th (2%), Doctorate (1%), 5-6th (1%), 1-4th (1%), Preschool (1%)						0
	Relationship	Husband (41%), Not-in-family (26%), Own child (16%), Unmarried (11%), Wife (4%), Other relative (2%)						0
	Race	White (85%), Black (10%), Asian / Pacific Islander (3%), American Indian / Eskimo (1%), Other (1%)						0
	Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%) Separated (3%), Widowed (2%), Married-AF-spouse (1%), Married-spouse-absent (1%)						0
	Occupation	15 categories						1843
	Country	42 categories: USA (90%)						583
Binomials	Salary [Label]	<=$50K (76%), >$50K (24%)						0
	Gender	Male (67%), Female (33%)						0
Real		Mean	Median	Std Dev	Skewness	Kurtosis	Range	
	Age	38.58	37	13.64	0.56	2.83	17 - 90	0
	Hours worked per week	40.44	40	12.35	0.23	5.92	1 - 99	0
	Education Number	10.08	10	2.57	-0.31	3.62	1 - 16	0
	Captial Gain	1078	0	7385	11.95	157.77	0 - 99999	0
	Capital Loss	87.3	0	403	4.59	23.37	0 - 4356	0
	Survey Weight	189778	178356	105550	1.45	9.22	12285 - 1484705	0

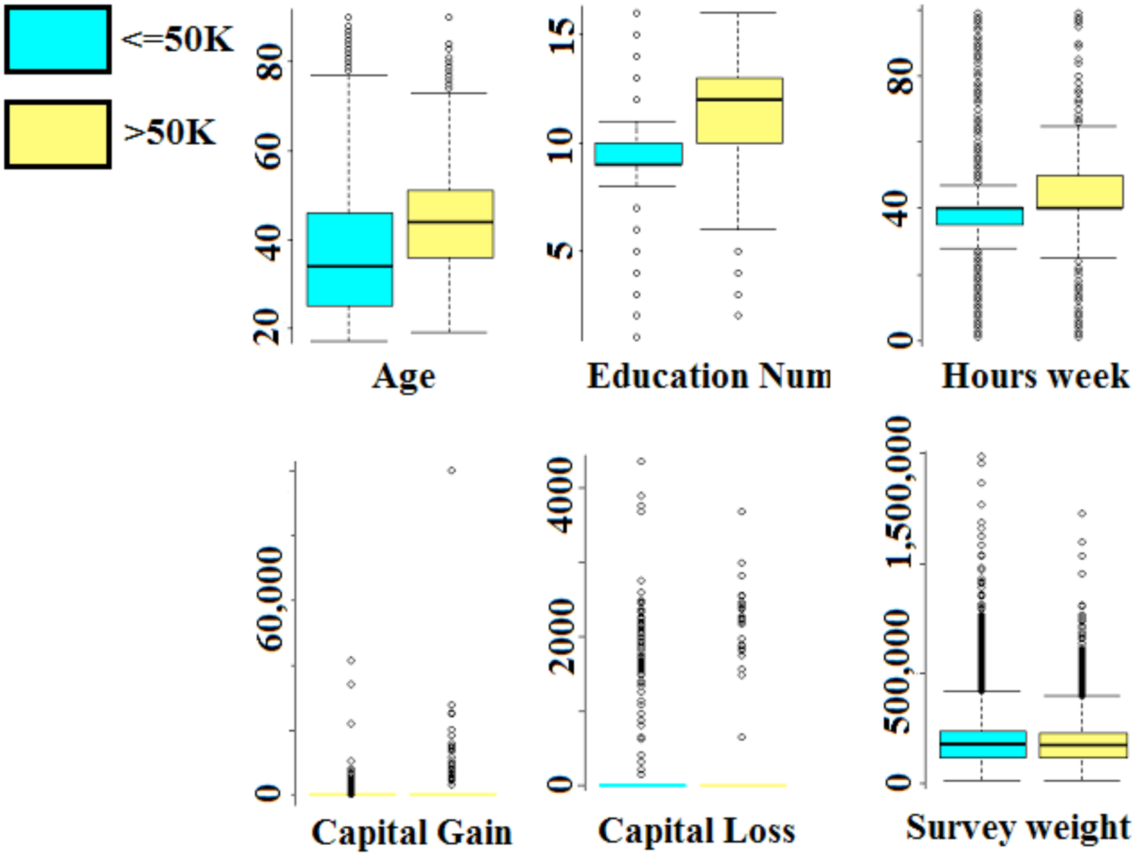
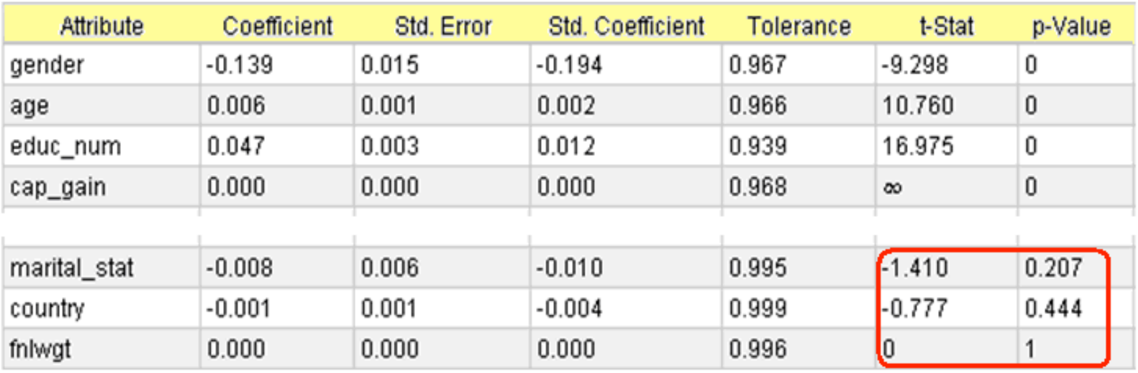


Figure 1. Boxplots of numeric attributes

The numeric attributes appear to contain a significant quantity of unique values (Table 2) and in the case of the survey weight attribute there were twenty one thousand unique values out of thirty one thousand instances, which may suggest that this attribute may not be significantly predictive. This was confirmed when a regression model was applied to the dataset where fnlwgt had a t-Statistic of zero and a p-Value of one (Figure 2). The other attributes were found to be reasonably significant with the exception of the country attribute (which contains a value of 'United-States' in ninety percent of instances) and the race attribute (which contains a value of 'White' in eighty five percent of instances), and may be candidates for removal during data preparation, and the marital status attribute which is explored further later in this report.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
gender	-0.139	0.015	-0.194	0.967	-9.298	0
age	0.006	0.001	0.002	0.966	10.760	0
educ_num	0.047	0.003	0.012	0.939	16.975	0
cap_gain	0.000	0.000	0.000	0.968	∞	0
marital_stat	-0.008	0.006	-0.010	0.995	-1.410	0.207
country	-0.001	0.001	-0.004	0.999	-0.777	0.444
fnlwgt	0.000	0.000	0.000	0.996	0	1

Figure 2. Regression classifier output on training dataset.

Table 2. Numerical attribute quantiles, unique values and outliers

	Statistics						Unique Values	Outliers
	Min	Q1	Median	Mean	Q3	Max		Limit [Qty outside limit]
age	17	28	37	38	48	90	73	0 [0] 78 [143]
survey wgt	12280	117800	178400	189800	237100	1485000	21648	0 [0] 425000 [865]
cap gain	0	0	0	1078	0	100000	119	
cap loss	0	0	0	87	0	4356	92	
hrs_week	1	40	40	40	45	99	94	32.5 [5516] 52.5 [3492]

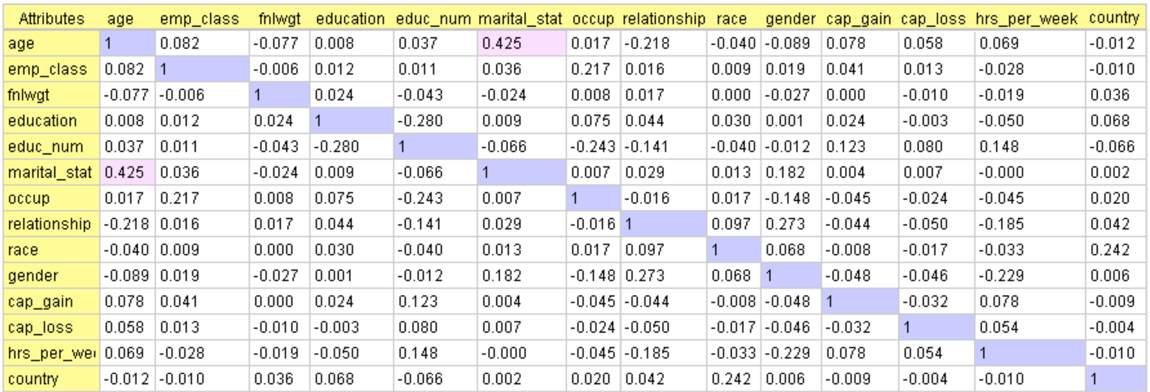
Although the capital gain and loss attributes have significant quantities of unique values (Table 2), the majority of the instances have a zero value with capital gain having ninety six percent zero values in the <50K class and seventy nine percent in the >50K class and capital loss having ninety eight percent in the <50K class and ninety percent in the >50K class. This may indicate that these attributes also may not be very predictive.

The numeric education number and nominal education level attributes were found to be fully correlated and therefore one of these attributes may be a good candidate for removal during modelling (Table 3). In general the other attributes were found to be weakly correlated (Figure 3). When the education number attribute was plotted for the class labels it was found that the lower values tend to predominate in the <50K class and higher levels in the >50K class which may indicate some predictive capability (Figure 4).

Table 3. Mapping between education number and education level attributes

Educ_num	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Education	Preschool	1st-4th	5th-6th	7th-8th	9th	10th	11th	12th	HS-grad	Some-college

Educ_num	*11*	*12*	*13*	*14*	*15*	*16*
Education	Assoc-voc	Assoc-acdm	Bachelors	Masters	Prof-school	Doctorate



Attributes	age	emp_class	fnlwgt	education	educ_num	marital_stat	occup	relationship	race	gender	cap_gain	cap_loss	hrs_per_week	country
age	1	0.082	-0.077	0.008	0.037	0.425	0.017	-0.218	-0.040	-0.089	0.078	0.058	0.069	-0.012
emp_class	0.082	1	-0.006	0.012	0.011	0.036	0.217	0.016	0.009	0.019	0.041	0.013	-0.028	-0.010
fnlwgt	-0.077	-0.006	1	0.024	-0.043	-0.024	0.008	0.017	0.000	-0.027	0.000	-0.010	-0.019	0.036
education	0.008	0.012	0.024	1	-0.280	0.009	0.075	0.044	0.030	0.001	0.024	-0.003	-0.050	0.068
educ_num	0.037	0.011	-0.043	-0.280	1	-0.066	-0.243	-0.141	-0.040	-0.012	0.123	0.080	0.148	-0.066
marital_stat	0.425	0.036	-0.024	0.009	-0.066	1	0.007	0.029	0.013	0.182	0.004	0.007	-0.000	0.002
occup	0.017	0.217	0.008	0.075	-0.243	0.007	1	-0.016	0.017	-0.148	-0.045	-0.024	-0.045	0.020
relationship	-0.218	0.016	0.017	0.044	-0.141	0.029	-0.016	1	0.097	0.273	-0.044	-0.050	-0.185	0.042
race	-0.040	0.009	0.000	0.030	-0.040	0.013	0.017	0.097	1	0.068	-0.008	-0.017	-0.033	0.242
gender	-0.089	0.019	-0.027	0.001	-0.012	0.182	-0.148	0.273	0.068	1	-0.048	-0.046	-0.229	0.006
cap_gain	0.078	0.041	0.000	0.024	0.123	0.004	-0.045	-0.044	-0.008	-0.048	1	-0.032	0.078	-0.009
cap_loss	0.058	0.013	-0.010	-0.003	0.080	0.007	-0.024	-0.050	-0.017	-0.046	-0.032	1	0.054	-0.004
hrs_per_wei	0.069	-0.028	-0.019	-0.050	0.148	-0.000	-0.045	-0.185	-0.033	-0.229	0.078	0.054	1	-0.010
country	-0.012	-0.010	0.036	0.068	-0.066	0.002	0.020	0.042	0.242	0.006	-0.009	-0.004	-0.010	1

Figure 3. Correlation matrix for the training dataset.

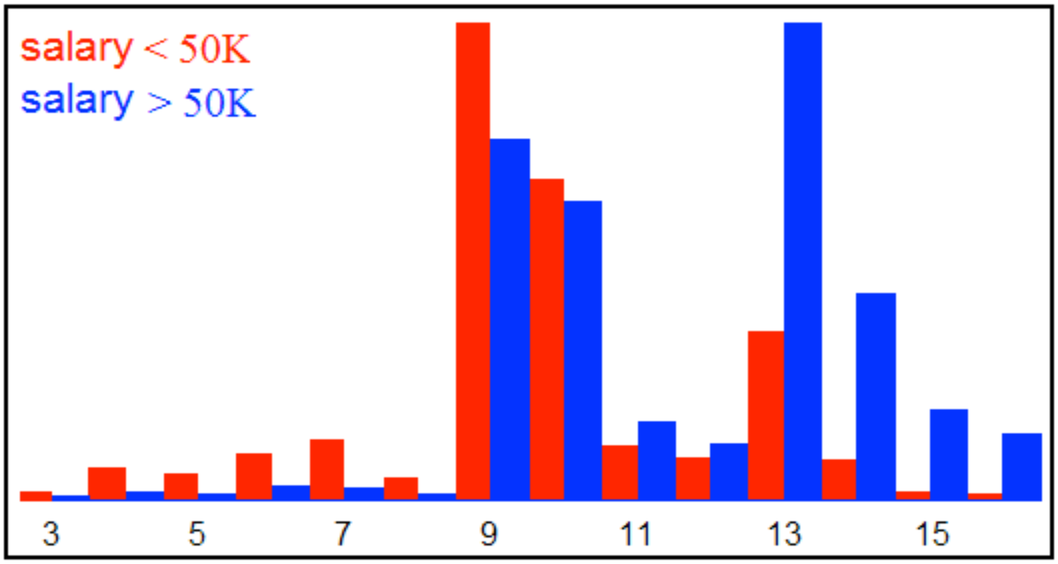


Figure 4. Normalised histogram of education level in class labels.

A slight bias was detected in the dataset where instances with a 'female' gender value have lower range of age values than instances with a 'male' gender value (Figure 5A). This may skew the predictive capability of the gender attribute to some degree as age values within >50K class instances tend to be higher (Figure 5B) than <50K instances. There is also a slight imbalance in gender with sixty seven percent of instances having a male value (Table 1).

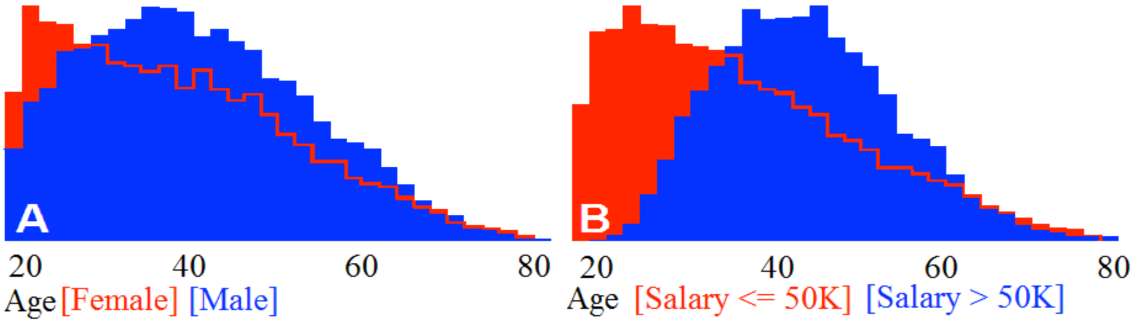
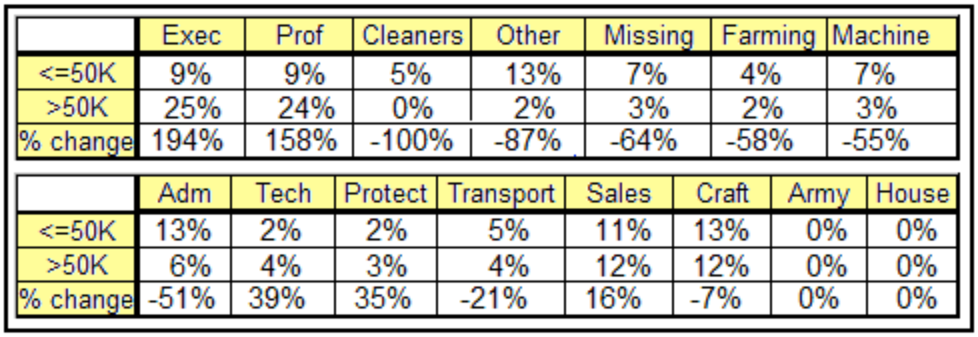


Figure 5. Histograms of age profile for gender and salary attributes

The occupation attribute was found to have some very infrequently occurring values such as Armed-Forces and Priv-house-serv, some values such as Cleaners and Other which are more highly correlated with the <50K class and other values such as Exec-managerial and Prof-specialty being more highly correlated with the <50K class (Table 4). This correlation between occupation categories and the class label (in particular the >50K label) was also observed when a decision tree classifier was applied to the training dataset (Figure 6) which suggests that this attribute may have a relatively high level of predictive capability.

Table 4. Comparison of occupation attribute values within the salary classes. A positive percentage change indicates an increased proportion in the >50K class.



	Exec	Prof	Cleaners	Other	Missing	Farming	Machine
<=50K	9%	9%	5%	13%	7%	4%	7%
>50K	25%	24%	0%	2%	3%	2%	3%
% change	194%	158%	-100%	-87%	-64%	-58%	-55%

	Adm	Tech	Protect	Transport	Sales	Craft	Army	House
<=50K	13%	2%	2%	5%	11%	13%	0%	0%
>50K	6%	4%	3%	4%	12%	12%	0%	0%
% change	-51%	39%	35%	-21%	16%	-7%	0%	0%

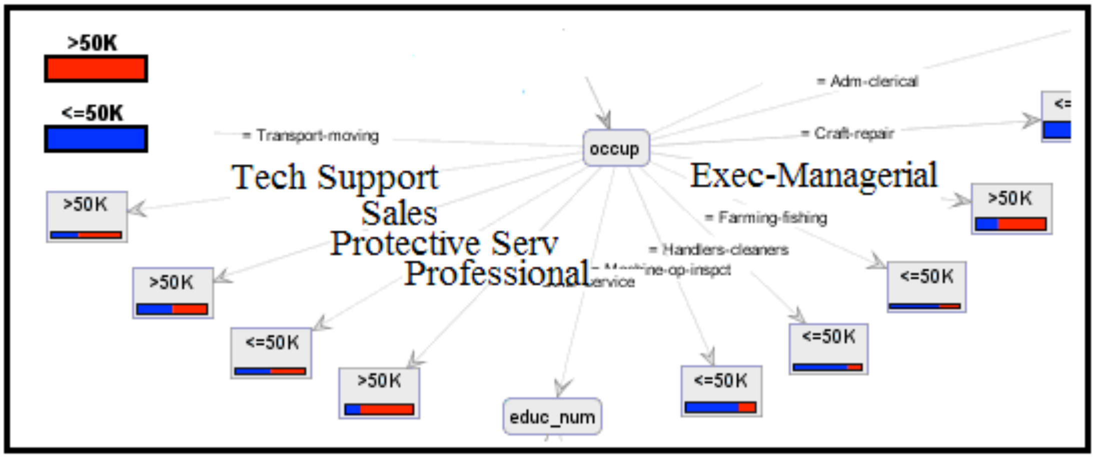
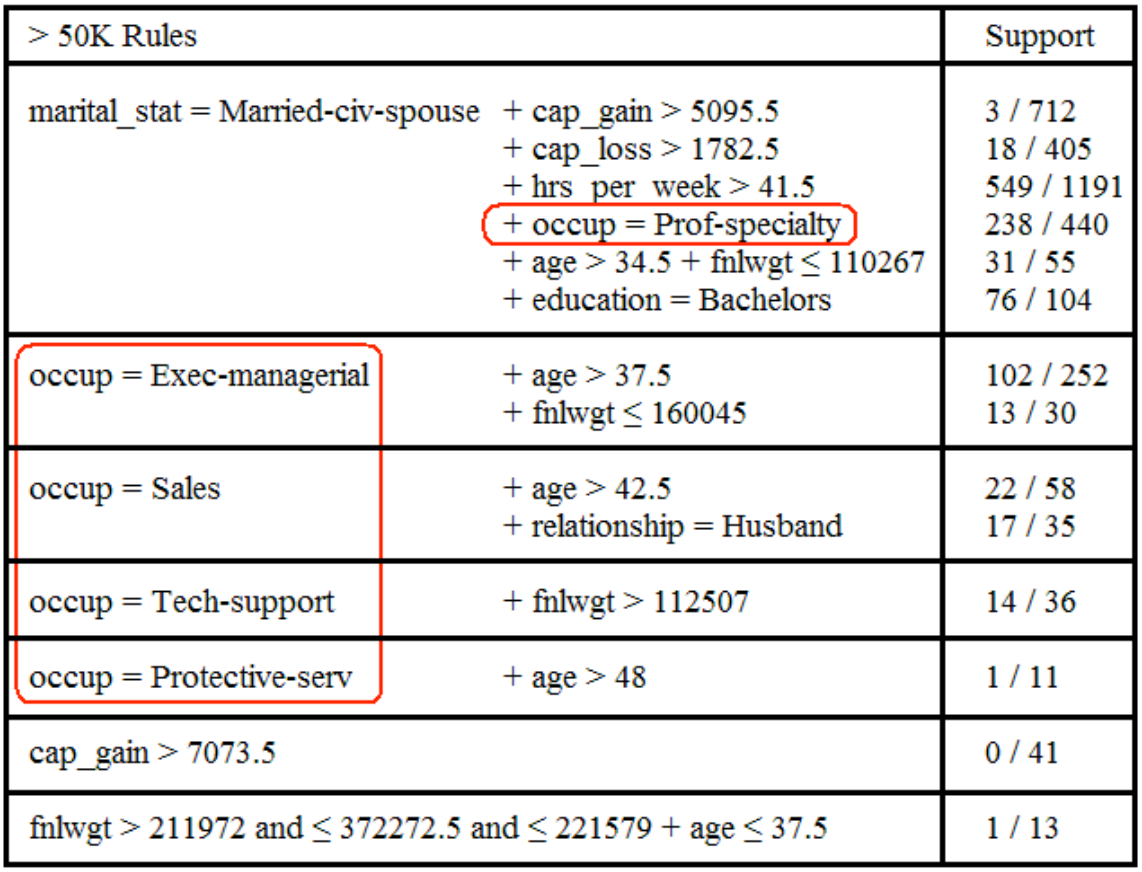


Figure 6. Correlation of occupation categories with >50K class from decision tree on training dataset.

In order to test the observations outlined above a rule induction classifier was applied to the training dataset which achieved an accuracy of 82.85%. The runtime was quite long at three hours and thirty minutes so this classifier would not be the optimal choice on this dataset unless it could be proven to significantly outperform other classifiers. The generated rules were very useful however in confirming some of the observations already noted such as the correlation between the occupation attribute and the >50K salary class (Table 5).

Table 5. Rules discovered for >50K salary class on training dataset.



> 50K Rules		Support
marital_stat = Married-civ-spouse	+ cap_gain > 5095.5 + cap_loss > 1782.5 + hrs_per_week > 41.5 + occup = Prof-specialty + age > 34.5 + fnlwgt ≤ 110267 + education = Bachelors	3 / 712 18 / 405 549 / 1191 238 / 440 31 / 55 76 / 104
occup = Exec-managerial	+ age > 37.5 + fnlwgt ≤ 160045	102 / 252 13 / 30
occup = Sales	+ age > 42.5 + relationship = Husband	22 / 58 17 / 35
occup = Tech-support	+ fnlwgt > 112507	14 / 36
occup = Protective-serv	+ age > 48	1 / 11
cap_gain > 7073.5		0 / 41
fnlwgt > 211972 and ≤ 372272.5 and ≤ 221579 + age ≤ 37.5		1 / 13

Verify data quality

In general it was found that the incidence of extreme outliers in the dataset was low with the exception of the hours_per_week (worked) attribute which has a significant percentage of outliers in both tails. The age and fnlwgt (survey weight) attributes had outliers in less than three percent of instances (Table 2). Rapidminer's Detect Outlier (Distances) operator (with a k value of six based on kNN classifier modelling) detected outliers primarily in the fnlwgt (survey weight) and capital gain attributes (Figure 7).

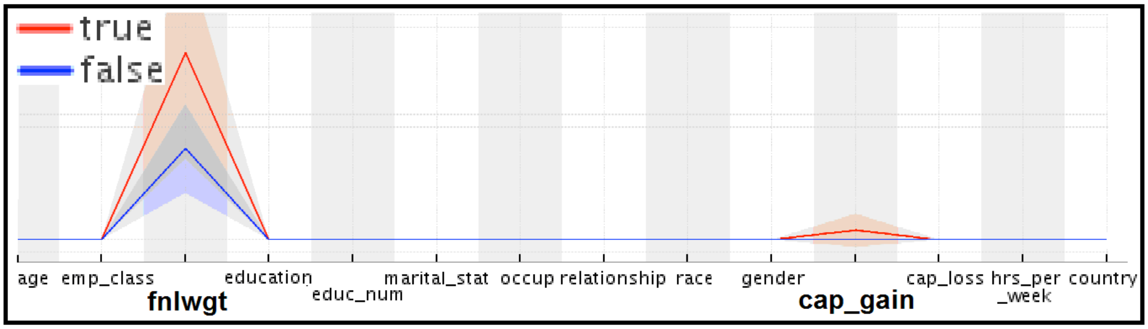


Figure 7. Deviation plot of the mean and range of values in outlier (true) and non-outlier (false) instances based on a mixed euclidean measures.

There are three attributes with missing values with an incidence rate of two percent for the country attribute and four percent for employment class and occupation (Table 1). It may be possible to impute the missing country values as this has a 'United-States' value ninety percent of the time and for employment class with 'Private' occurring in seventy three percent of instances. Imputing values for the occupation attribute may be more challenging as this attribute's values are more evenly distributed. As the overall incidence rate of missing values is quite low at less than five percent, it may be possible to remove instances with missing values without the need for value imputation or replacement.

The dataset was also checked for inconsistencies and conflicts such as male gender with a relationship value of wife or unmarried marital status with a relationship value of husband/wife etc, but very few instances of this type of issue were found indicating that the data is of a reasonable quality. With thirty two thousand available instances (or thirty thousand if all missing value instances were removed) and low levels of data duplication (only twenty four duplicate instances exist in the training dataset) there should be a sufficient quantity of clean data available to learn a classification model. There is some class imbalance within the dataset as seventy six percent of instances are in the <50K class, but the eight thousand instances in the >50K class should be sufficient to detect the patterns within this class or otherwise boosting can be applied.

4. Data Preparation

Select Data

In order to select a classifier for data preparation, a ROC curve was generated for multiple classifiers including Naïve Bayes, Rule Induction, kNN and Decision Tree with Naïve Bayes achieving the highest AUC on the unmodified training dataset (Figure 8). These classifiers had been selected as suitable candidates as the dataset has a mix of nominal and numeric attributes with a binomial label.

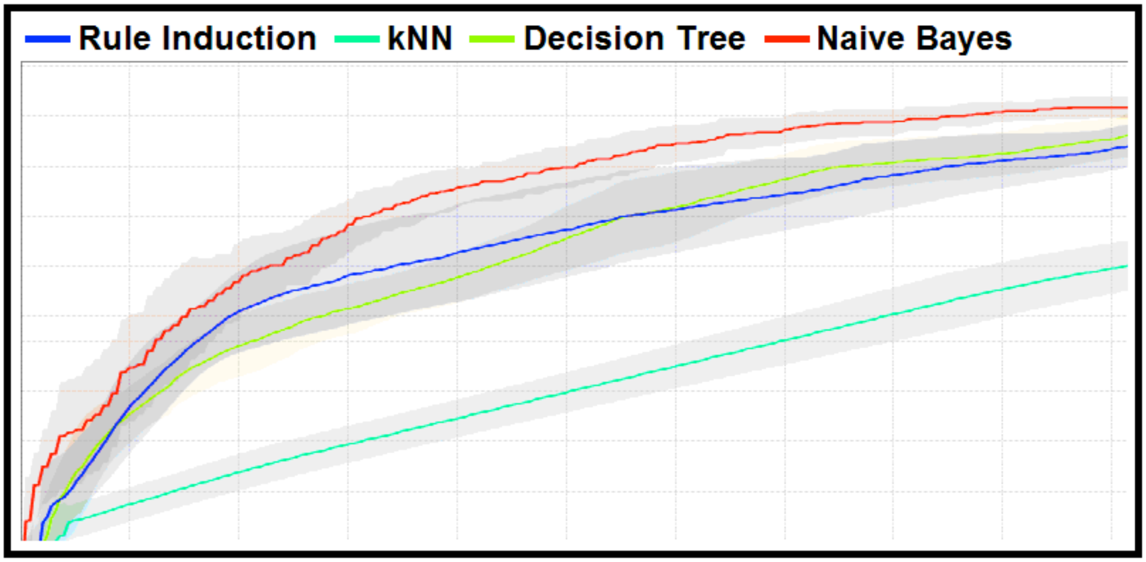
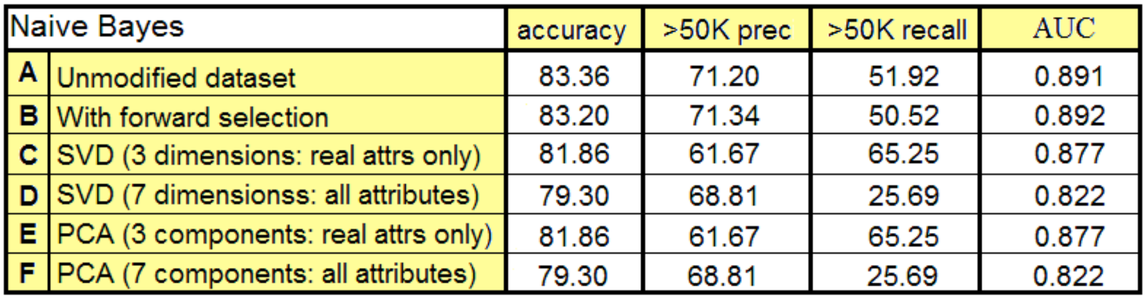


Figure 8. Initial ROC curves on training dataset.

To investigate the applicability of feature reduction to the full training dataset forward selection was applied to a Naïve Bayes classifier (Table 6 A). The results of this test indicate that although the >50K class precision was improved slightly over Naïve Bayes (without forward selection) the overall performance was degraded slightly (Table 6 B). The application of forward selection did however provide useful information as the fnlwgt (survey weight) and country attributes had been dropped without significantly impacting the performance of the Naïve Bayes classifier confirming the earlier assertion during data exploration that these attributes may not be significantly predictive.

Singular value decomposition (SVD) and principle component analysis (PCA) were then applied to the dataset in two ways. Initially SVD and PCA (with an optimal quantity of three dimensions/ components) were applied to the numerical attributes only which were joined to the nominal attributes and passed to a Naïve Bayes classifier with both achieving an eighty two percent accuracy (Table 6 C/E). SVD and PCA (with an optimal quantity of six dimensions/components) were then applied to all attributes, where the nominal attributes had been mapped to real values, and both approaches now achieved an accuracy of seventy nine percent (Table 6 D/F). These results indicate that the above approaches to feature reduction on the unmodified training dataset do not appear to be useful in improving classifier accuracy.

Table 6. Naïve Bayes with forward selection, SVD and PCA on training dataset.



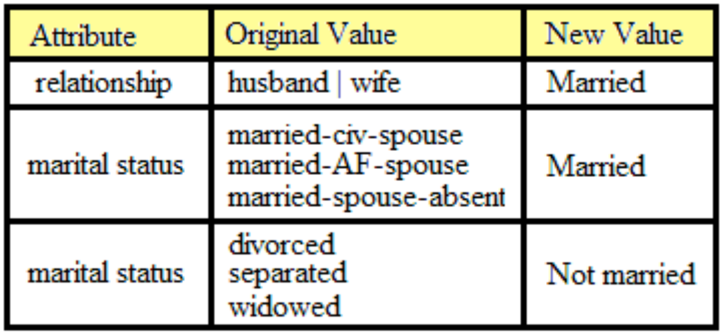
Naive Bayes		accuracy	>50K prec	>50K recall	AUC
A	Unmodified dataset	83.36	71.20	51.92	0.891
B	With forward selection	83.20	71.34	50.52	0.892
C	SVD (3 dimensions: real attrs only)	81.86	61.67	65.25	0.877
D	SVD (7 dimensionss: all attributes)	79.30	68.81	25.69	0.822
E	PCA (3 components: real attrs only)	81.86	61.67	65.25	0.877
F	PCA (7 components: all attributes)	79.30	68.81	25.69	0.822

Construct Data

When the marital status and relationship attribute values were modified slightly to reduce the number of categories (Table 7) it was observed that the information in the marital status attribute could be inferred to some degree from the 'unmarried' category in the relationship attribute suggesting that marital status might be a candidate for removal.

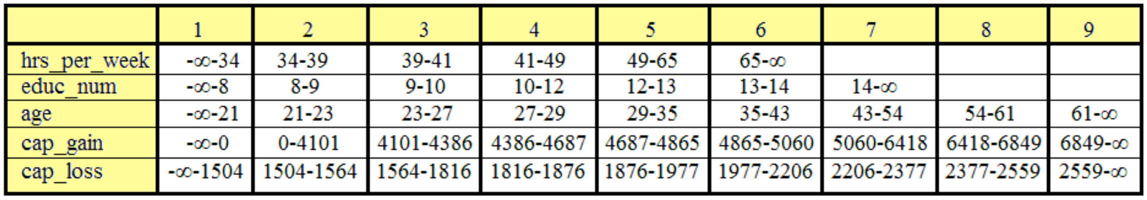
Discretisation was applied to the numerical attributes to determine if the performance of the Naïve Bayes classifier could be improved. Initially entropy binning was investigated on the training dataset but a performance of eighty two percent was poorer than that achieved without binning. However the output (Table 8) from this exercise did provide a useful starting point in suggesting bin boundaries for the following discretisation work.

Table 7. Relationship and marital status attribute transformations.



Attribute	Original Value	New Value
relationship	husband | wife	Married
marital status	married-civ-spouse married-AF-spouse married-spouse-absent	Married
marital status	divorced separated widowed	Not married

Table 8. Entropy binning output on training dataset.

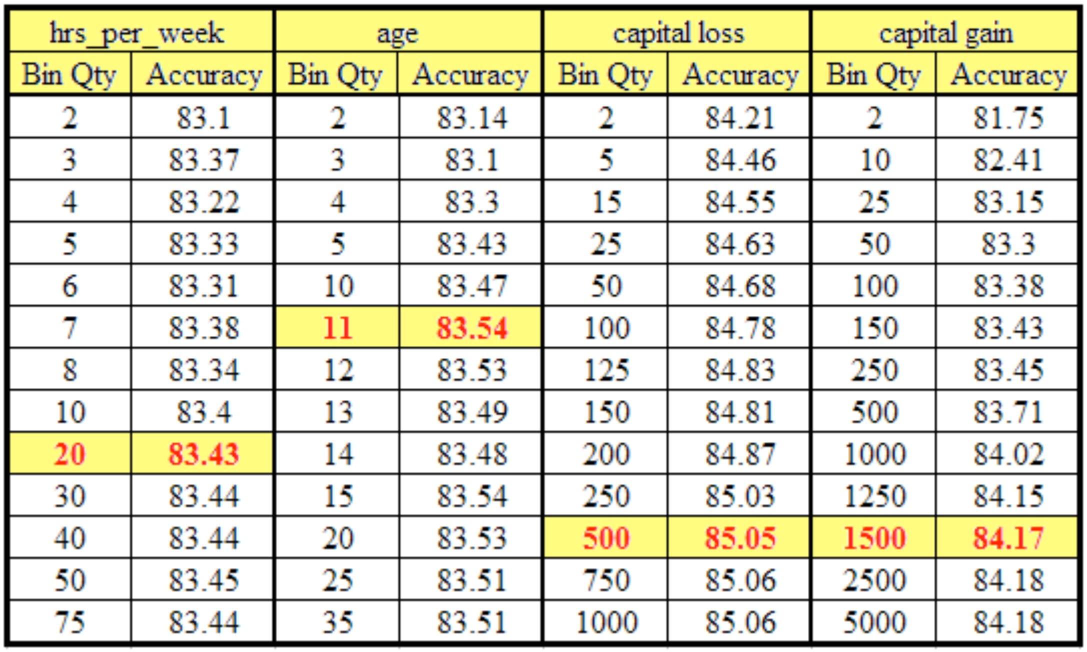


	1	2	3	4	5	6	7	8	9
hrs_per_week	-∞-34	34-39	39-41	41-49	49-65	65-∞			
educ_num	-∞-8	8-9	9-10	10-12	12-13	13-14	14-∞		
age	-∞-21	21-23	23-27	27-29	29-35	35-43	43-54	54-61	61-∞
cap_gain	-∞-0	0-4101	4101-4386	4386-4687	4687-4865	4865-5060	5060-6418	6418-6849	6849-∞
cap_loss	-∞-1504	1504-1564	1564-1816	1816-1876	1876-1977	1977-2206	2206-2377	2377-2559	2559-∞

By iteratively testing various bin boundary combinations using Rapidminer's Optimise Parameter operator an optimal bin quantity for the hrs_per_week (worked) attribute was found to be twenty with the most highly populated bin (by a factor of ten) having bin boundaries of thirty five and forty with a mean of thirty nine. It was found that using a two bin approach with a bin boundary of this mean value of thirty nine proved to be equally effective. (Table 9).

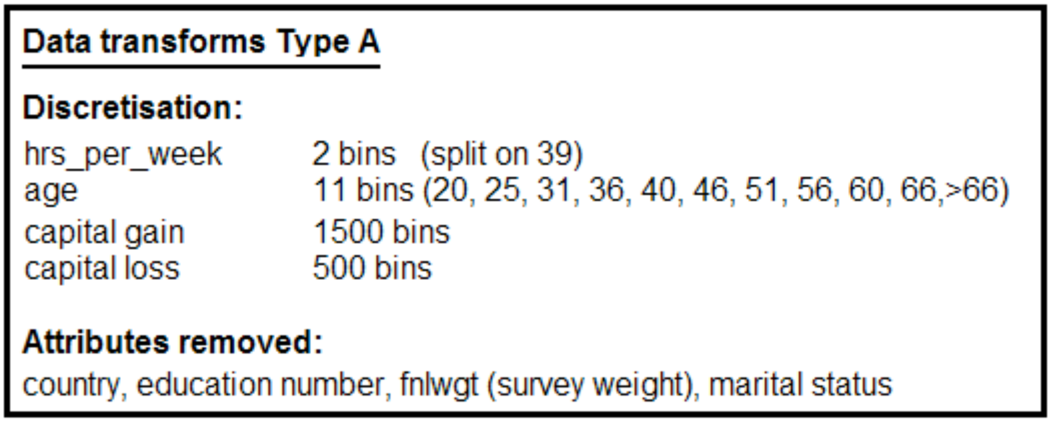
Using the binning operator in rapidminer the optimal bin quantity for the age attribute was found to be eleven (Table 9) and based on the generated bin boundaries and work with R to determine the data distribution twenty bins were selected at boundaries ranging from twenty to sixty six (specifically 20, 25, 31, 36, 40, 46, 51, 56, 60, 66 and over 66) which slightly improved the classifier's performance. The more variate capital gain and loss attributes were found to have optimal binning quantities of five hundred and fifteen hundred respectively as increasing the bin quantity beyond this points did not improve the model's performance significantly (Table 9).

Table 9. Numerical attribute binning results



hrs_per_week		age		capital loss		capital gain	
Bin Qty	Accuracy	Bin Qty	Accuracy	Bin Qty	Accuracy	Bin Qty	Accuracy
2	83.1	2	83.14	2	84.21	2	81.75
3	83.37	3	83.1	5	84.46	10	82.41
4	83.22	4	83.3	15	84.55	25	83.15
5	83.33	5	83.43	25	84.63	50	83.3
6	83.31	10	83.47	50	84.68	100	83.38
7	83.38	11	83.54	100	84.78	150	83.43
8	83.34	12	83.53	125	84.83	250	83.45
10	83.4	13	83.49	150	84.81	500	83.71
20	83.43	14	83.48	200	84.87	1000	84.02
30	83.44	15	83.54	250	85.03	1250	84.15
40	83.44	20	83.53	500	85.05	1500	84.17
50	83.45	25	83.51	750	85.06	2500	84.18
75	83.44	35	83.51	1000	85.06	5000	84.18

During initial modelling on the training dataset with Naïve Bayes it was found that removing the country (which has the same value in ninety percent of cases), education number (which is correlated with education level), survey weight (which is highly variate) and marital status (which appears to contain similar information to the relationship attribute as outlined earlier) attributes did not affect the model's performance. Therefore these four attributes will be considered for removal during future modelling. The results of these initial data transformations are summarised below and will be referred to in this report as data transformations type A (Figure 9).



Data transforms Type A

Discretisation:

hrs_per_week	2 bins (split on 39)
age	11 bins (20, 25, 31, 36, 40, 46, 51, 56, 60, 66,>66)
capital gain	1500 bins
capital loss	500 bins

Attributes removed:

country, education number, fnlwgt (survey weight), marital status

Figure 9. Data transformations type A

5. Modelling

Select modelling technique

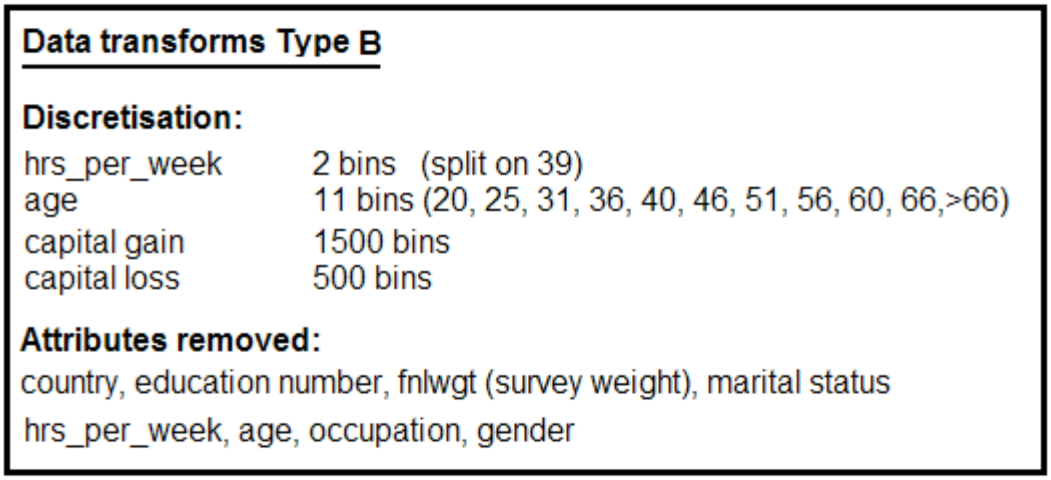
As stated above as the dataset has mixed numerical and nominal attributes with a binomial class label a Decision Tree, kNN, Naïve Bayes and Rule Induction classifier had been selected for initial modelling.

Generate Test Design

During modelling the training dataset will initially be used to evaluate each classifier's performance relative to that achieved by Naïve Bayes on the unmodified training dataset (Table 6 A) and the optimal classifiers will then be evaluated on the test dataset in terms of overall performance and ability to support the primary business objective of maximising the return on investment.

Build and Assess the model

As described in the data preparation section above applying forward selection on the unmodified training dataset with Naïve Bayes had not been very successful (Table 6) it was decided during modelling to revisit this idea but this time the data transformations (Type A) above were applied ahead of forward selection. This approach proved to be quite successful with a Naïve Bayes performance improvement of three percent (Table 10). Examination of the model's example-set showed that forward selection on the transformed data had also additionally removed the hrs_per_week, age, occupation and gender attributes which will be referred to as data transformations type B in this report (Figure 10).



Data transforms Type B

Discretisation:

hrs_per_week	2 bins (split on 39)
age	11 bins (20, 25, 31, 36, 40, 46, 51, 56, 60, 66,>66)
capital gain	1500 bins
capital loss	500 bins

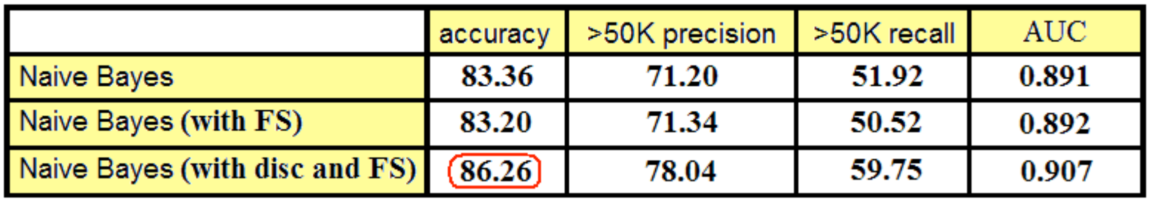
Attributes removed:

country, education number, fnlwgt (survey weight), marital status

hrs_per_week, age, occupation, gender

Figure 10. Data transformations type B

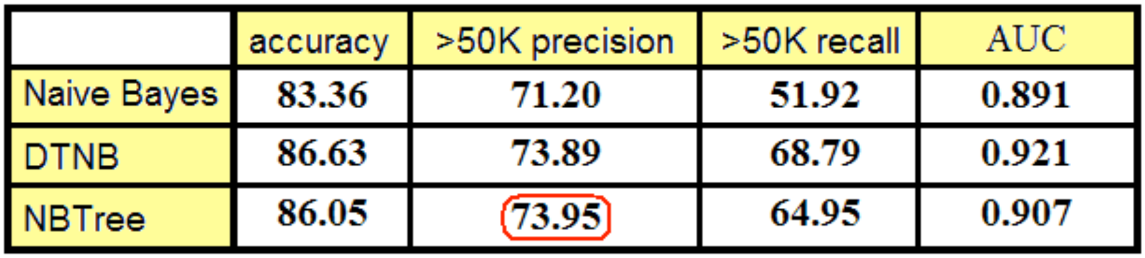
Table 10. Naïve Bayes performance on the training dataset with unmodified data, data with forward selection only and with forward selection to the transformed data.



	accuracy	>50K precision	>50K recall	AUC
Naive Bayes	83.36	71.20	51.92	0.891
Naive Bayes (with FS)	83.20	71.34	50.52	0.892
Naive Bayes (with disc and FS)	86.26	78.04	59.75	0.907

As additional attribute transformations did not further improve the performance of the Naïve Bayes classifier the WEKA extension was imported into Rapidminer in order to access some hybrid tree classifiers. As Naïve Bayes had performed well on the training dataset the NBTree and DTNB classifiers appeared to be good candidates as they both combine Naïve Bayes with a decision tree which may provide useful insights into the data even if further performance improvements are not achieved. DTNB is a cost-sensitive learner which attempts to minimize mis-classification costs using a single Naïve Bayes classifier in the background to assist in making globally influenced branching decisions at the nodes. The NBTree classifier is slightly different in that it creates a Naïve Bayes classifier at each node in order to determine how the node should be split. The DTNB and NBTree did indeed prove to be the best performing classifiers in the Rapidminer WEKA classifier set on the unmodified training dataset and out-performed Naïve Bayes (Table 11 and Figure 11).

Table 11. Naïve Bayes and hybrid classifier performances on unmodified training dataset



	accuracy	>50K precision	>50K recall	AUC
Naive Bayes	83.36	71.20	51.92	0.891
DTNB	86.63	73.89	68.79	0.921
NBTree	86.05	73.95	64.95	0.907

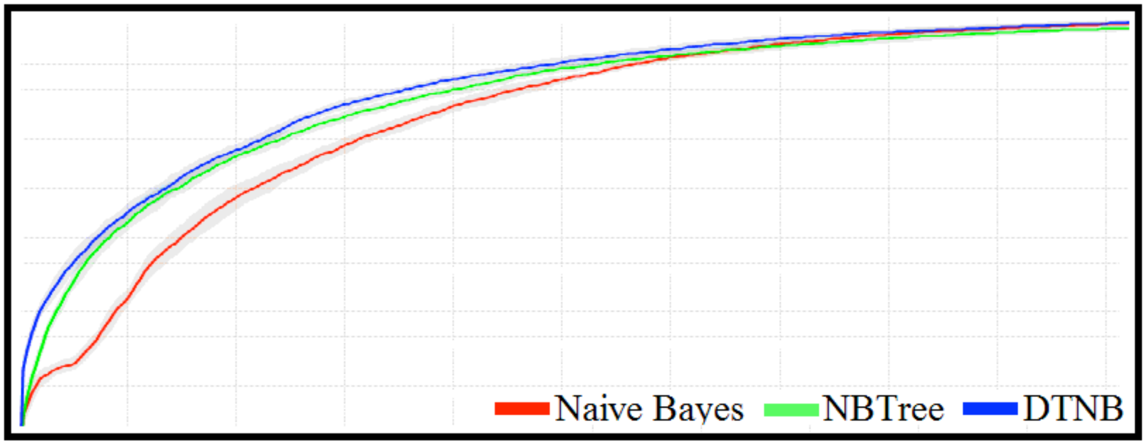
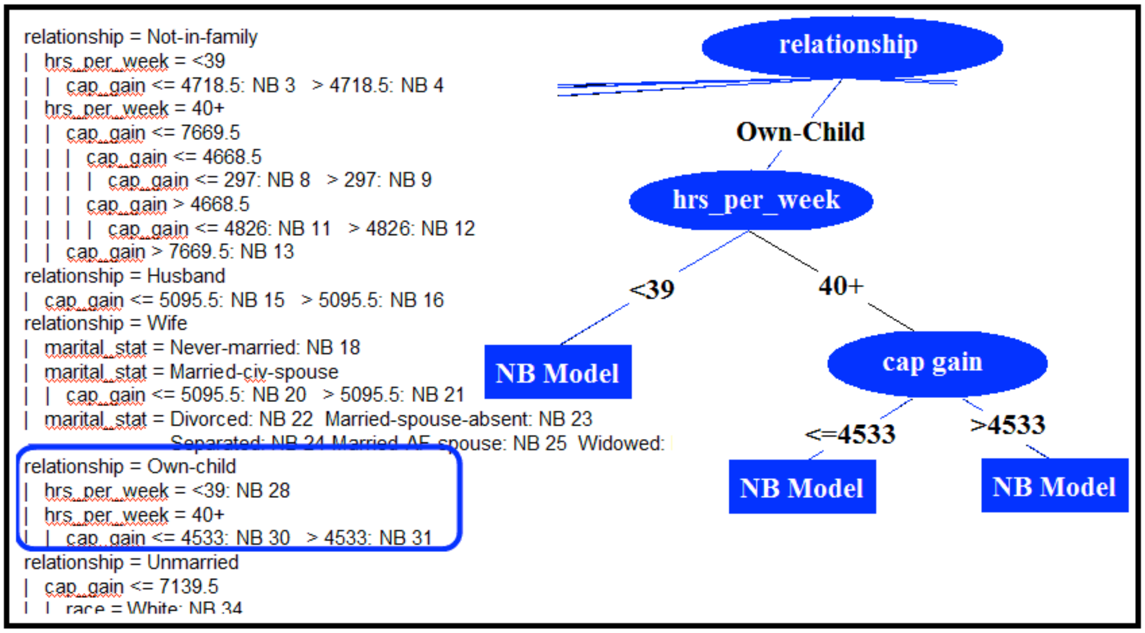


Figure 11. ROC curves on unmodified training dataset.

The rules generated in the NBTree output (Figure 12) were quite useful in determining attribute significance which was broadly in agreement with what had been discovered using other classifiers.

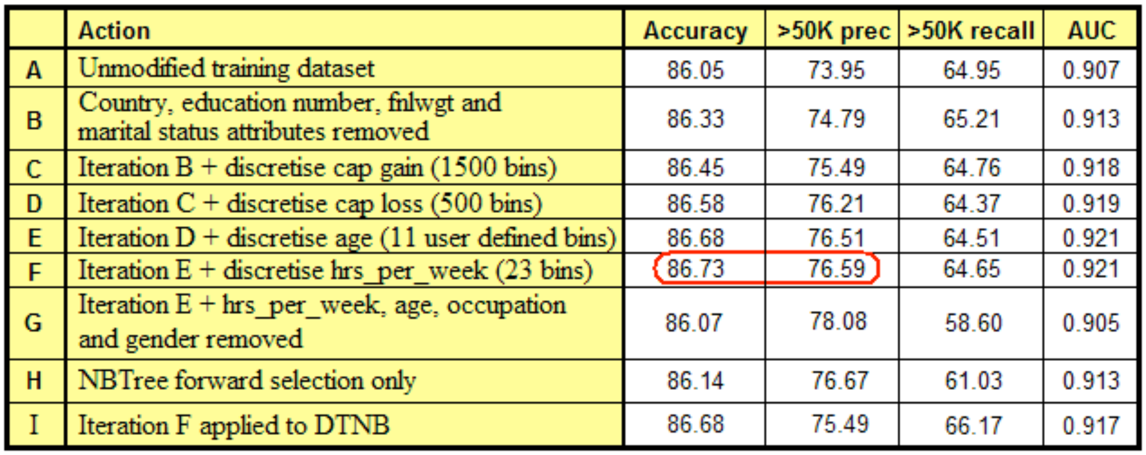


```
relationship = Not-in-family
|   hrs_per_week = <39
|   |   cap_gain <= 4718.5: NB 3   > 4718.5: NB 4
|   hrs_per_week = 40+
|   |   cap_gain <= 7669.5
|   |   |   cap_gain <= 4668.5
|   |   |   |   cap_gain <= 297: NB 8   > 297: NB 9
|   |   |   cap_gain > 4668.5
|   |   |   |   cap_gain <= 4826: NB 11   > 4826: NB 12
|   |   cap_gain > 7669.5: NB 13
relationship = Husband
|   cap_gain <= 5095.5: NB 15   > 5095.5: NB 16
relationship = Wife
|   marital_stat = Never-married: NB 18
|   marital_stat = Married-civ-spouse
|   |   cap_gain <= 5095.5: NB 20   > 5095.5: NB 21
|   marital_stat = Divorced: NB 22  Married-spouse-absent: NB 23
                 Separated: NB 24 Married-AF-spouse: NB 25  Widowed:
relationship = Own-child
|   hrs_per_week = <39: NB 28
|   hrs_per_week = 40+
|   |   cap_gain <= 4533: NB 30   > 4533: NB 31
relationship = Unmarried
|   cap_gain <= 7139.5
|   |   race = White: NB 34
```

Figure 12. NBTree classifier output on training dataset. The portion of the tree shown reflects the rules indicated in the highlighted area where each leaf contains a local implementation of a Naïve Bayes classifier.

The data transformations outlined above were then applied to the NBTree classifier on training dataset with the type A data transformations achieving 86.73% (Table 12 F) and the type B transformations achieving a slightly lower accuracy of 86.07% (Table 12 G). For comparison purposes forward selection with no data transformations was then applied to NBTree which had an accuracy of 86.14% (Table 12 H) which falls within the range of accuracies achieved with data transformation type A. Finally the type A transformations were applied to the DTNB classifier which achieved a similar accuracy to NBTree of 86.68% (Table 12 I).

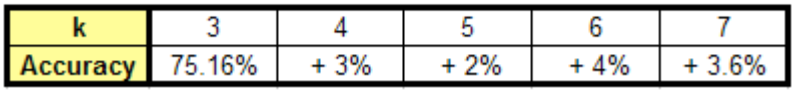
Table 12. Discretisation applied to NBTree classifier on training dataset.



	Action	Accuracy	>50K prec	>50K recall	AUC
A	Unmodified training dataset	86.05	73.95	64.95	0.907
B	Country, education number, fnlwgt and marital status attributes removed	86.33	74.79	65.21	0.913
C	Iteration B + discretise cap gain (1500 bins)	86.45	75.49	64.76	0.918
D	Iteration C + discretise cap loss (500 bins)	86.58	76.21	64.37	0.919
E	Iteration D + discretise age (11 user defined bins)	86.68	76.51	64.51	0.921
F	Iteration E + discretise hrs_per_week (23 bins)	86.73	76.59	64.65	0.921
G	Iteration E + hrs_per_week, age, occupation and gender removed	86.07	78.08	58.60	0.905
H	NBTree forward selection only	86.14	76.67	61.03	0.913
I	Iteration F applied to DTNB	86.68	75.49	66.17	0.917

Although kNN had performed quite poorly on the unmodified dataset (Figure 8) it was decided (as per the CRISP-DM framework) to return to the data preparation stage and apply the data transformations outlined above to kNN. The first step was the selection of an optimal value for k which was found to be six based on the classifier's accuracy on a sample of the training dataset (Table 13).

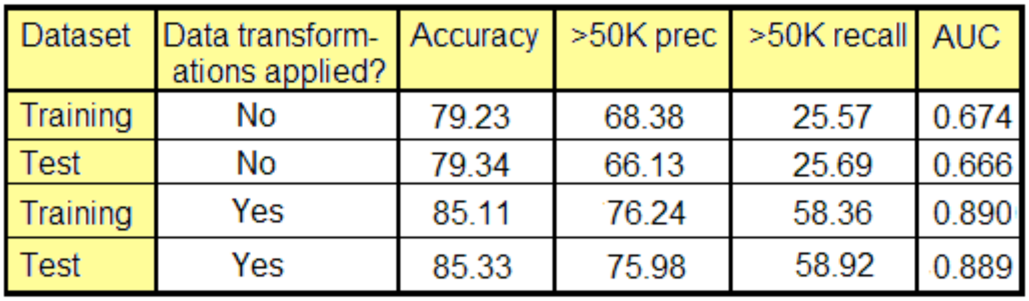
Table 13. kNN performance for k values from 3 to 7 on unmodified training dataset.



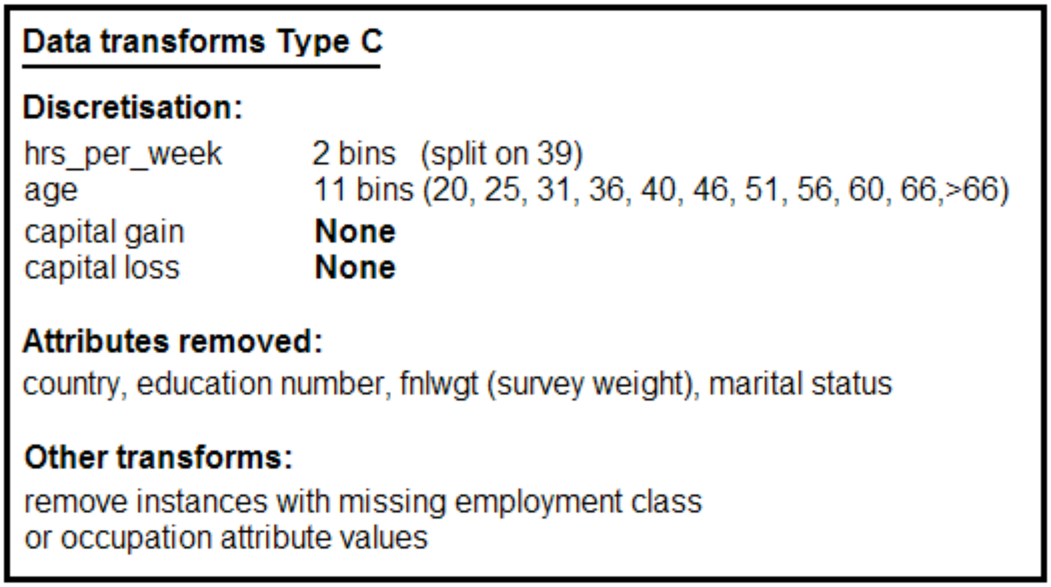
k	3	4	5	6	7
Accuracy	75.16%	+ 3%	+ 2%	+ 4%	+ 3.6%

It was found that the type A transformations required some modification to achieve an optimal performance with kNN, where the discretisation on capital gain and loss was removed (as kNN preferred this lower level of data granularity) and missing values were removed, which were not an issue for Naïve Bayes. This improved kNN's performance from 79% to 85% (Table 14) and provided further independent validation of the usefulness of the proposed data transformations. The data transformations applied to kNN will be referred to as data transformations type C (Figure 13).

Table 14. kNN (k=6) classifier performance.



Dataset	Data transform-ations applied?	Accuracy	>50K prec	>50K recall	AUC
Training	No	79.23	68.38	25.57	0.674
Test	No	79.34	66.13	25.69	0.666
Training	Yes	85.11	76.24	58.36	0.890
Test	Yes	85.33	75.98	58.92	0.889



Data transforms Type C

Discretisation:

hrs_per_week	2 bins (split on 39)
age	11 bins (20, 25, 31, 36, 40, 46, 51, 56, 60, 66,>66)
capital gain	**None**
capital loss	**None**

Attributes removed:

country, education number, fnlwgt (survey weight), marital status

Other transforms:

remove instances with missing employment class
or occupation attribute values

Figure 13. Data transformations type C

The ROC curve also indicated a much improved performance with kNN with the type C transformation as it now lies between the NBTree and Naïve Bayes ROC curves (Figure 14). The original ROC curve for kNN on the unmodified training dataset (kNN(Original) has also been included in the ROC curves for comparison purposes with an AUC little better than 0.5.

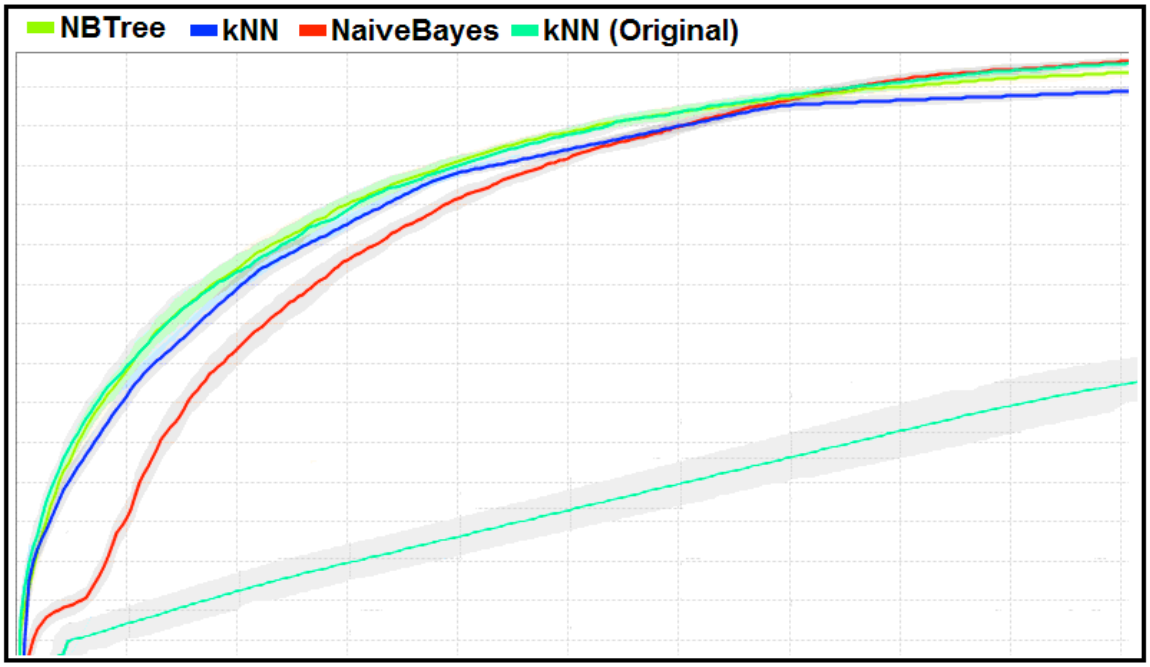


Figure 14 ROC curves of NBTree, Naïve Bayes and kNN (for both transformed and unmodified training data).

In order to evaluate the performance of a cost sensitive classifier a MetaCost operator was then applied to the training dataset. The cost matrix contained a value of one hundred and twenty five in the lower left which is the cost per offer and a value of minus five hundred in the lower right which is the potential maximum revenue per correct classification. It was found that an optimal performance was achieved when MetaCost was used with a logistic model tree (LMT). An LMT classifier replaces sub-trees with linear regression functions if the data subset at that node is suitable for this type of classification. A goal of LMT is to avoid overfitting as we work down into smaller subsets of data within the tree and to simplify the overall structure of the tree without impacting performance. This model worked quite well on the training dataset with a performance of 80% and applying the data transformations outlined above did not significantly improve performance. This classifier had a runtime of over four hours on the training dataset which means that a significant performance gain on other classifiers would be required to justify it's selection which was not the case here. This classifier was also applied to the test dataset as described later in this report.

One interesting observation with the MetaCost(LMT) classifier was that the cost attribute in the output appeared to form two clusters (Figure 15) where the lower cluster would contain instances with a higher probability of being in the >50K salary class.

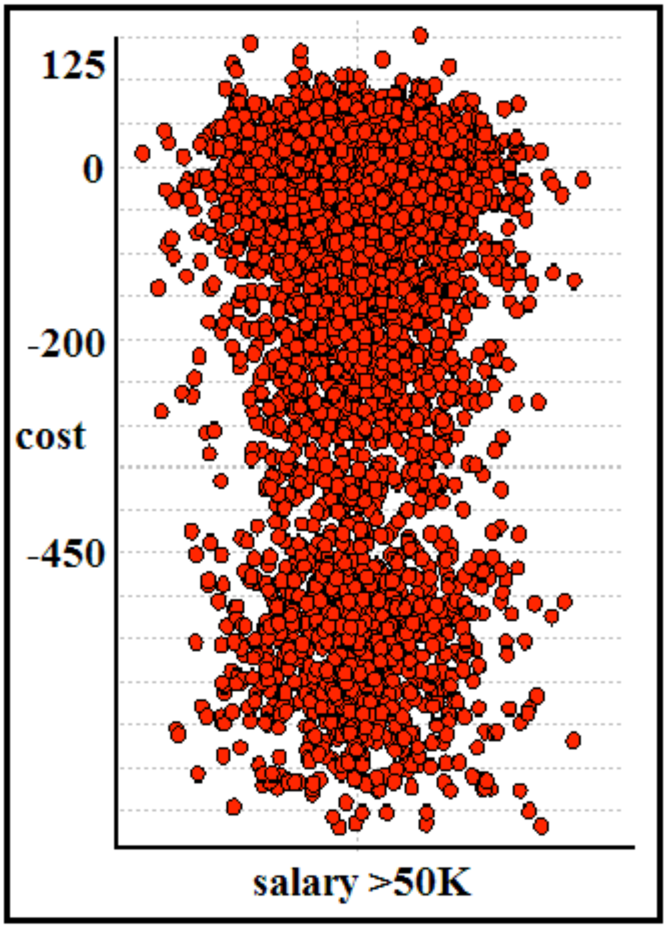
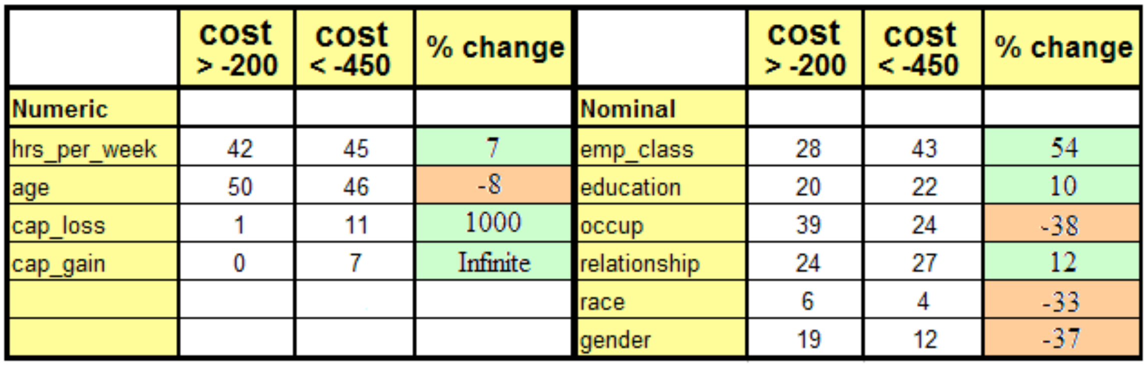


Figure 15. Scatter plot of MetaCost output cost attribute for >50K label instances

To explore this further the cluster attribute means were calculated for instances with cost values greater than minus two hundred and for instances with cost values less (more negative) than minus four hundred and fifty (Table 15). The differences in the attribute means indicated that the capital gain and loss attributes were significantly correlated (as demonstrated by the mean deltas) with the >50K class label. Further work in this area could involve mapping the nominal attributes to meaningful real values to determine the degree of mean differences between the class labels but that work is beyond the scope of this report.

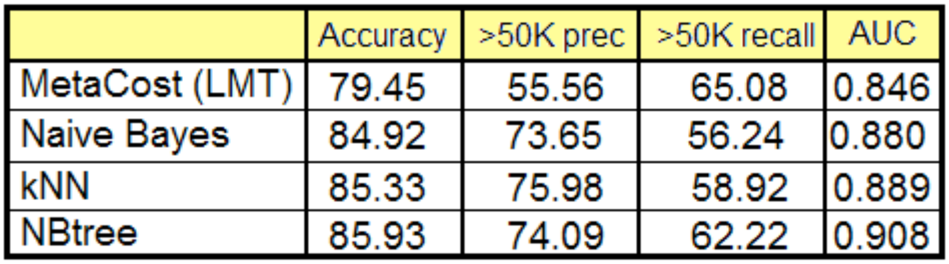
Table 15. Mean values of >50K salary class cost attribute clusters.



	cost > -200	cost < -450	% change		cost > -200	cost < -450	% change
Numeric				Nominal			
hrs_per_week	42	45	7	emp_class	28	43	54
age	50	46	-8	education	20	22	10
cap_loss	1	11	1000	occup	39	24	-38
cap_gain	0	7	Infinite	relationship	24	27	12
				race	6	4	-33
				gender	19	12	-37

The four best performing classifiers found during the modelling phase (Naïve Bayes, NBTree, MetaCost(LMT) and kNN) were then evaluated on the test dataset with the appropriate data transformations as outlined above. It was found that NBTree had the best overall performance (Table 16). Naïve Bayes had the lowest runtime at one minute, followed by NBTree at three minutes, kNN at fifteen minutes and the MetaCost(LMT) at three hours.

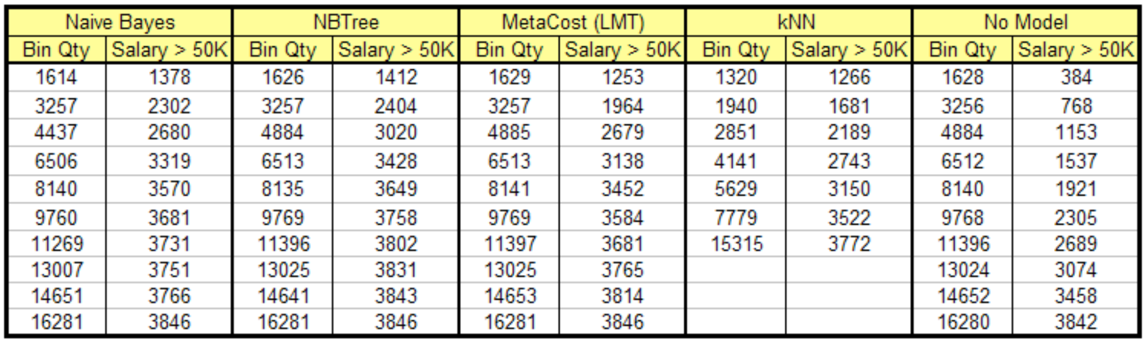
Table 16. Classifier performance on test dataset.



	Accuracy	>50K prec	>50K recall	AUC
MetaCost (LMT)	79.45	55.56	65.08	0.846
Naive Bayes	84.92	73.65	56.24	0.880
kNN	85.33	75.98	58.92	0.889
NBtree	85.93	74.09	62.22	0.908

The lift for each classifier was then calculated on the test dataset in order to evaluate relative performances with respect to the business objective of maximising return on investment and overall profitability (Table 17).

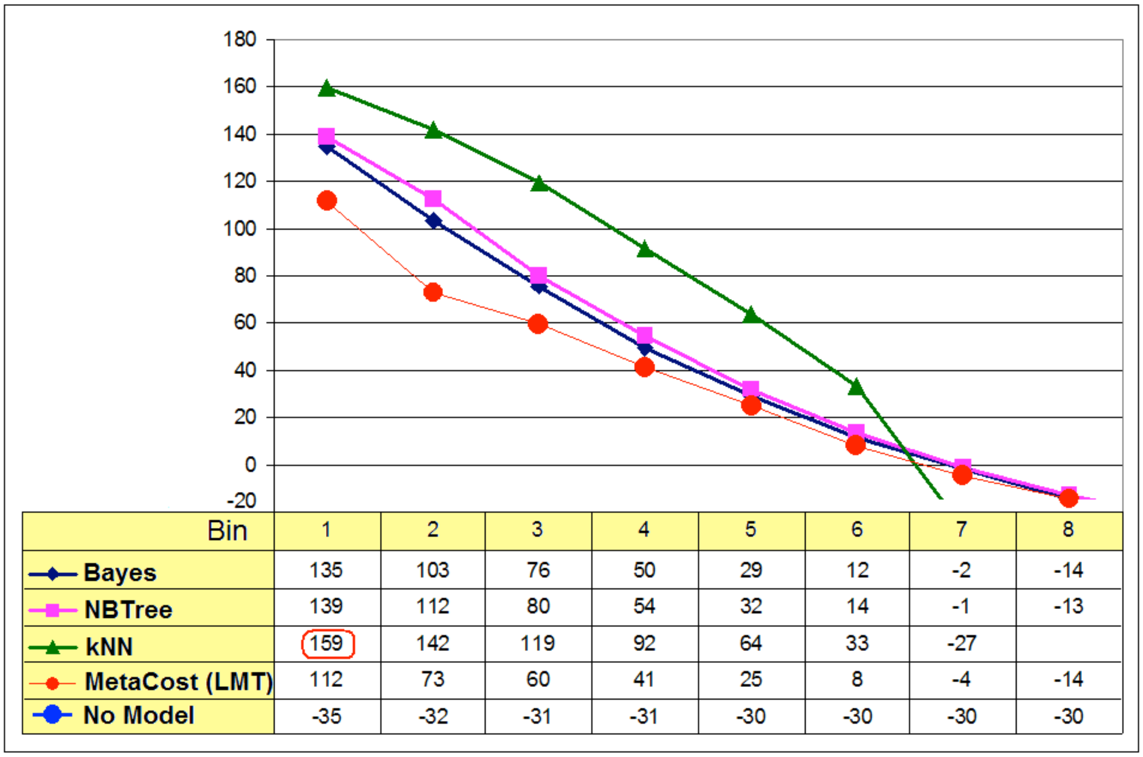
Table 17. Classifier lift numbers on the test dataset. 'Bin Qty' is the total number of prospects in each of the ten bins and the 'Salary>50K' quantity is the expected number of correctly classified >50K instances.



Naive Bayes		NBTree		MetaCost (LMT)		kNN		No Model	
Bin Qty	Salary > 50K	Bin Qty	Salary > 50K	Bin Qty	Salary > 50K	Bin Qty	Salary > 50K	Bin Qty	Salary > 50K
1614	1378	1626	1412	1629	1253	1320	1266	1628	384
3257	2302	3257	2404	3257	1964	1940	1681	3256	768
4437	2680	4884	3020	4885	2679	2851	2189	4884	1153
6506	3319	6513	3428	6513	3138	4141	2743	6512	1537
8140	3570	8135	3649	8141	3452	5629	3150	8140	1921
9760	3681	9769	3758	9769	3584	7779	3522	9768	2305
11269	3731	11396	3802	11397	3681	15315	3772	11396	2689
13007	3751	13025	3831	13025	3765			13024	3074
14651	3766	14641	3843	14653	3814			14652	3458
16281	3846	16281	3846	16281	3846			16280	3842

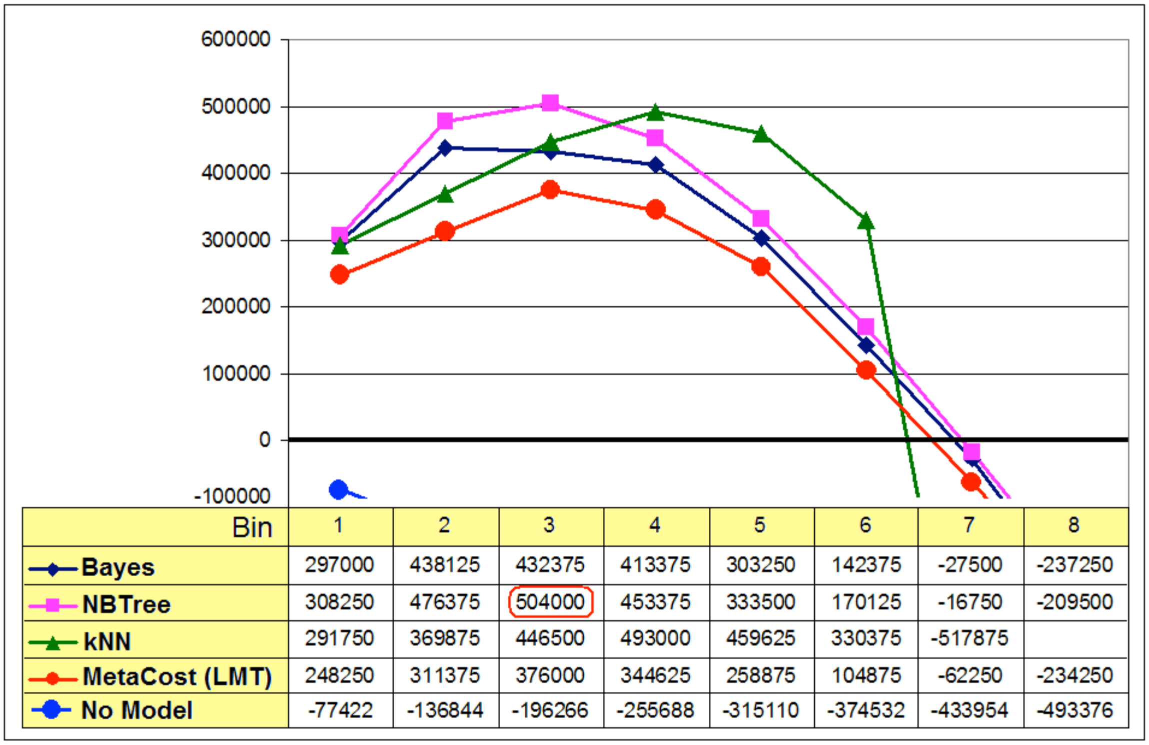
Based on this lift data the overall return on investment (Figure 16) and profitability (Figure 17) were calculated per bin for each model (Appendix A contains a worked example of these calculations).

The kNN classifier (with type C data transformations) consistently returned the highest return on investment with a maximum value of at one hundred and fifty nine when offers were sent to prospects in the first lift bin only. The NBTree model (with type A data transformations) was the most profitable in real terms with offers sent to prospects in the first three bins generating an expected revenue of one million one hundred thousand dollars for an initial investment of six hundred and twenty thousand dollars. These calculations are based on the relatively small test dataset which was used in this project and would be expected to scale up when applied to larger datasets. For comparison purposes the profitability and return on investment was calculated for each bin in a 'No Model' scenario based on the assumption that each bin contains twenty four percent >50K instances (which reflects the degree of occurrence in the general population) which had a negative return on investment and was a loss making exercise regardless of the number of prospects contacted.



Bin	1	2	3	4	5	6	7	8
Bayes	135	103	76	50	29	12	-2	-14
NBTree	139	112	80	54	32	14	-1	-13
kNN	159	142	119	92	64	33	-27	
MetaCost (LMT)	112	73	60	41	25	8	-4	-14
No Model	-35	-32	-31	-31	-30	-30	-30	-30

Figure 16. Classifier return on investment per bin on test dataset.



Bin	1	2	3	4	5	6	7	8
Bayes	297000	438125	432375	413375	303250	142375	-27500	-237250
NBTree	308250	476375	504000	453375	333500	170125	-16750	-209500
kNN	291750	369875	446500	493000	459625	330375	-517875	
MetaCost (LMT)	248250	311375	376000	344625	258875	104875	-62250	-234250
No Model	-77422	-136844	-196266	-255688	-315110	-374532	-433954	-493376

Figure 17. Classifier profitability per bin on test dataset.

6. Evaluation

During the CRISP-DM data understanding phase of the project some of the attributes were found to be predominantly single valued such as the country (with 'United-States') and capital gain and loss (with 0) in over ninety percent of instances. The survey weight attribute had the opposite issue where there was a large number of unique values. Later work with forward selection on Naïve Bayes confirmed that the country and survey weight attributes could be removed successfully without impacting classifier performance. The numeric education number and nominal education level attributes were found to be fully correlated (with education number being removed later during data preparation), and a correlation was also discovered between the marital status and relationship attributes after some basic category aggregation had been carried out and also between categories in the occupation attribute and the >50K class label.

The overall data quality was quite good with a low occurrence of conflicting attribute values and with over thirty thousand clean instances in the training dataset and fifteen thousand in the training dataset (with both having low levels of data duplication) there was a sufficient quantity of clean variate data in both class labels to successfully create (with training data) and evaluate (with test data) the various classifiers. Some outliers were detected in the data but this was not found to seriously impact classifier performance as the percentage of affected instances was low in general with the hours_per_week (worked) attribute having slightly elevated levels but as binning was found to be beneficial on this attribute the impact of outliers was reduced. The percentage of records with missing values was quite low with country at two percent and employment class and occupation at four percent which did not affect the classification work except in the case of kNN where the affected instances were successfully removed. Some biases were detected in the data such as female (gender attribute) instances generally having a lower set of values than male instances but incidences of this type were not found to be significant.

The general observations made during data exploration were confirmed when an initial rule induction classifier was applied to the training data and again when the NBTree output was analysed. When the models which were deemed to be appropriate for this dataset (with mixed numerical and nominal attributes and a binomial label) were applied to the unmodified training data the Naïve Bayes classifier had the best performance.

During initial data preparation it was found that the optimal data transformation for Naïve Bayes involved discretising the hrs_per_week (worked), age, capital gain and capital loss attributes and removing the country, education number, survey weight and marital status attributes (which had been noted as possible candidates for removal earlier). This useful data transformation was tagged as type A in this report. When the type A data transformation was applied ahead of forward selection on a Naïve Bayes classifier a further performance improvement was achieved by additionally removing the hrs_per_week, age, occupation and gender attributes which was tagged as type B data transformation.

The type A data transformations were then applied to the hybrid DTNB and NBTree classifiers on the training dataset with improved performance on both and a slight modification to the type A (where the discretisation on the capital gain and loss attributes was removed as were instances with missing values) proved optimal for the kNN classifier which was tagged as type C data transforms.

The modelling work on the training dataset appears to indicate that there is a classifier accuracy limit of just below eighty seven percent which is supported by other work such as the Naïve Bayes performance given at eighty three to four percent with discretisation (Kaya, 2008) and also at eighty four percent (Kohavi, 1996) and NBTree results posted by UCI (UCI Archive, 2011) indicating a performance of eighty five percent. It therefore appears that the proposed models and associated data transforms outlined in this report are close to optimal.

During the modelling phase the training dataset had been used to evaluate the relative performance of each of the selected classifiers and these models were then evaluated on the test dataset on the level of accuracy and level of support (based on lift) for the primary goal of maximising return on investment. NBTree had the highest accuracy at 85.93% and was also the most profitable (on the current test dataset) when the first three (of ten) lift bins were used. The kNN classifier had the highest return on investment on all lift bins. The 'No Model' approach resulted in losses for all bins.

Based on these results it would seem that Naïve Bayes (with type B data transformations) offers good performance with minimum runtime and may be a good choice for the initial validation of new datasets. The maximum profit levels were generated by the NBTree classifier (with type A data transformations) as it reached more prospects than the other classifiers with a slightly longer runtime then Naïve Bayes. The maximum return on investment was generated by the kNN classifier (with type C data transformations) with a k value of 6 but this carried a runtime overhead which was five times longer than NBTree.

As the goals of identifying useful classifiers to support the business objective of maximising return on investment and the provision of descriptions of the attributes which were deemed to be significant have been achieved it is believed that the results outlined above satisfactorily support the stated business objectives and should form a good foundation for future classification work on similar datasets.

7. Appendix A

As an example of the calculations used to generate the return on investment and profitability charts we will work through the figures for the first bin quantities found with the Naïve Bayes model (Table 17). The calculations are based on an initial setup cost of eighteen thousand dollars, with an offer cost of one hundred and twenty five dollars and an acceptance revenue of five hundred dollars:

Naïve Bayes Bin 1:

Number of prospects in bin = 1614
Predicted number of salaries > 50K = 1378

Predicted number of acceptances = 1378 * 0.75 = 1033
(*seventy five percent of individuals with salaries exceeding fifty thousand dollars are expected to accept the current offer strategy*)

Total cost = setup cost + total cost of offers sent
= 18000 + (1614 * 125)
= 219,500

Total revenue = number of acceptances * revenue per acceptance
= 1033 * 500
= 516,500

Profit = total revenue - total cost
= 516500 - 219500
= **297,000**

ROI = (profit / total cost) * 100
= (297000/219500) * 100
= **135**

8. References

Kaya, F., 2008, 'Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers', *University of Maryland publications*.

Kohavi, R., 1996, 'Scaling Up the Accuracy of Naïve Bayes Classifiers: a DecisionTree Hybrid', Proceedings of the second international conference on knowledge discovery and data mining.

UCI Archive, 2011, *http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names*