भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT BANGALORE

Certificate Program in Business Analytics and Intelligence (BATCH 9)

Module 3: Predictive Analytics

(Assignment 2 – 120 Points)

Instructions

1. This is a **take-home** assignment. You are free to discuss the assignment questions with your classmates. However, you are not allowed to copy the answers from other students.
2. Answer all questions.
3. **Show all work and give adequate explanations to get credit. The mathematical equations of the solutions should be clearly mentioned.**
4. Encircle or underline your final answer for each part.
5. Follow the file name format before you upload the submission on Moodle **Module3_Assignment2_(Your Name).pdf.**
6. **Completed assignment MUST BE uploaded on Moodle by 8th October 2018. Assignments submitted after 8th October 2018 will not be accepted.**
7. 10 marks will be deducted over every week for any Late Submission.
8. Course Completion Certificate will not be awarded for Non-Submission of the assignments.

Question 1 (20 Points)

Value of 506 properties were analysed using the variables described in Table 1.a.

Table 1.a. Data Dictionary

S.No	Variable	Variable Type	Code in SPSS output
1	Value of property	Numerical (in lakhs of rupees)	Price
2	Crime Rate	Numerical	CRIM
3	Proportion of Residential Land	Numerical	RES
4	SEZ (Special Economic Zone)	Binary	1 = Property is close to SEZ; 0 otherwise
5	Average Number of Rooms per Dwelling	Numerical	RM
6	Age of the property	Numerical	Age
7	Weighted Distance to 5 major employment centers	Numerical	DIS
8	Index of Accessibility to Major Highways	Numerical	Highway

Descriptive Statistics are shown in Table 1.1:

Table 1.1 Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CRIM	506	.00632	88.97620	3.6135236	8.60154511
Price	506	5	50	22.53	9.197
AGE	506	2.9	100.0	68.575	28.1489
RES	506	.46	27.74	11.1368	6.86035
DIS	506	1.13	12.13	3.7950	2.10571
Valid N (listwise)	506				

Correlation Between variables are shown in Table 1.2

Table 1.2: Correlation Matrix

	CRIM	RES	SEZ	RM	AGE	DIS	Highway	Price
CRIM	1	.407	-.056	-.219	.353	-.380	.626	-.388
RES	.407	1	.063	-.392	.645	-.708	.595	-.484
SEZ	-.056	.063	1	.091	.087	-.099	-.007	.175
RM	-.219	-.392	.091	1	-.240	.205	-.210	.695
AGE	.353	.645	.087	-.240	1	-.748	.456	-.377
DIS	-.380	-.708	-.099	.205	-.748	1	-.495	.250
Highway	.626	.595	-.007	-.210	.456	-.495	1	-.382
Price	-.388	-.484	.175	.695	-.377	.250	-.382	1

<u>Model 1 :</u> Y (Price) = β0 + β1 x CRIM

SPSS model outputs are shown in Tables 1.3 and 1.4.

Table 1.3 Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1		.		8.484

a. Predictors: (Constant), CRIM

b. Dependent Variable: Price

Table 1.4 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	24.033	.409		58.740	.000
	CRIM		.044		-9.460	.000

a. Dependent Variable: Price

Question 1.1 (2 points)

Calculate the rate at which the property price changes for unit change in the Crime Rate (Rate).

Question 1.2 (2 points)

Francis Galton, researcher at Flat Dekho.Com claims that for every unit increase in crime rate, the price will decrease by at least INR 30,000. Check whether Galton is correct at 95% confidence level.

Question 1.3 (2 Points)

Can we claim that when CRIM = 0, the average price of the property will be 24.033. Clearly state your arguments.

Question 1.4 (3 Points)

What is the maximum value of price at 95% confidence interval when CRIM = 1?

A second model is developed between Price and the binary variable SEZ. The outputs are shown in Tables 1.5 and 1.6 and the residual plots are shown in Figures 1.1 and 1.2.

Table 1.5 Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.175[a]	.031	.029	9.064

a. Predictors: (Constant), SEZ
b. Dependent Variable: Price

Table 1.6 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	22.094	.418		52.902	.000
	SEZ	6.346	1.588	.175	3.996	.000

a. Dependent Variable: Price

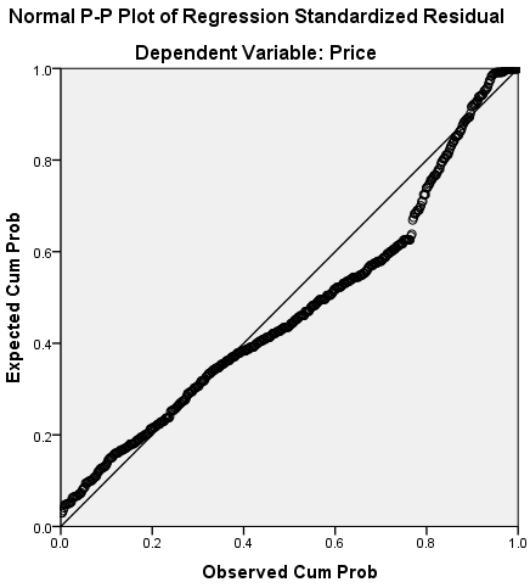


Figure 1.1 Normal Probability Plot

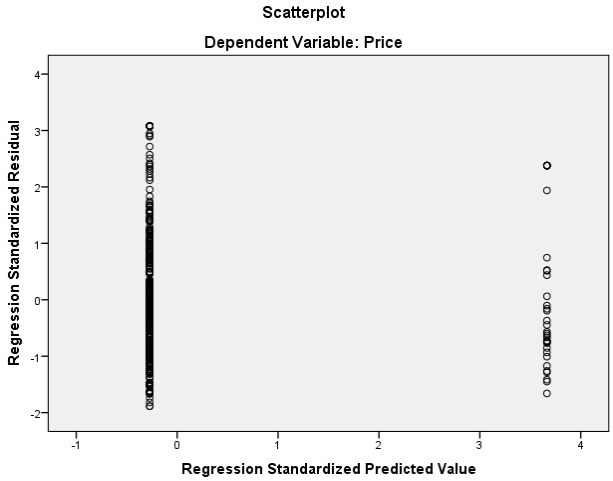


Figure 1.2. Plot of Standardized predicted value Vs Standardized Residuals

Question 1.5 (2 Points)

What is the probability that a property near SEZ will be at least 40 lakhs?

Question 1.6 (3 points)

a) Is there an evidence for heteroscedasticity in model 2?
b) What can you conclude from the probability plot in Figure 1?
c) What are the implications on the model based on your answer to question (a) and (b)?

A stepwise regression model is developed using outputs are shown in Tables 1.7 and 1.8.

Table 1.7 Model Summary[g]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.695[a]	.484	.483	6.616
2	.736[b]	.542	.540	6.237
3				
4	.762[d]	.581	.578	5.977
5	.767[e]	.589	.585	5.927
6	.782[f]	.612	.607	5.765

a. Predictors: (Constant), RM

b. Predictors: (Constant), RM, CRIM

c. Predictors: (Constant), RM, CRIM, RES

d. Predictors: (Constant), RM, CRIM, RES, SEZ

e. Predictors: (Constant), RM, CRIM, RES, SEZ, Highway

f. Predictors: (Constant), RM, CRIM, RES, SEZ, Highway, AGE

g. Dependent Variable: PRICE

Table 1.8 Stepwise Regression Output

Model		Unstandardized Coefficients		Standardized Coefficients	T	Correlations		
		B	Std. Error	Beta		Zero-order	Partial	Part
1	(Constant)	-34.671	2.650		-13.084			
	RM	9.102	.419	.695	21.722	.695	.695	.695
2	(Constant)	-29.245	2.588		-11.300			
	RM	8.391	.405	.641	20.726	.695	.679	.625
	CRIM	-.265	.033	-.248	-8.011	-.388	-.336	-.242
3	(Constant)	-22.141	2.867		-7.722			
	RM	7.648	.420	.584	18.227	.695	.631	.536
	CRIM	-.201	.035	-.188	-5.811	-.388	-.251	-.171
	RES	-.239	.046	-.179	-5.217	-.484	-.227	-.153
4	(Constant)	-20.863	2.834		-7.362			
	RM	7.433	.416	.568	17.888	.695	.624	.517
	CRIM	-.189	.034	-.176	-5.542	-.388	-.240	-.160
	RES	-.265	.045	-.198	-5.820	-.484	-.252	-.168
	SEZ	4.563	1.061	.126	4.300	.175	.189	.124
5	(Constant)	-16.458	3.155		-5.217			

	RM	7.289	.415	.557	17.575	.695	.618	.504
	CRIM	-.205	.034	-.192	-6.003	-.388	-.259	-.172
	RES	-.382	.059	-.285	-6.463	-.484	-.278	-.185
	SEZ	4.310	1.056	.119	4.083	.175	.180	.117
	Highway	-.557	.181	-.127	-3.073	.250	-.136	-.088
6	(Constant)	-8.993	3.363		-2.674			
	RM	7.182	.404		17.782	.695	.623	.496
	CRIM	-.194	.033		-5.822	-.388	-.252	-.162
	RES	-.318	.059	-.238	-5.420	-.484	-.236	-.151
	SEZ	4.499	1.027	.124	4.379	.175	.192	.122
	Highway	-1.154	.208	-.264	-5.555	.250	-.241	-.155
	AGE	-.077	.014		-5.430	-.377	-.236	-.151

Question 1.7 (2 Points)

What is the value of R-square at step 3 of the stepwise regression output in Table 1.7?

Question 1.8 (2 Points)

What is the possible reason for reduction in the value of the coefficient for the Variable “RM” between Model 1 and Model 2 of the Stepwise Regression?

Question 1.9 (2 points)

Which variable has highest impact on the price of the property?

Question 2 (20 Points)

Box office collection of 150 Bollywood movies were analysed using the variables described in Table 2.1.

Table 2.1. Data Dictionary

S.No	Variable	Variable Type	Code in SPSS output
1	Box office Collection (Y)	Numerical (in crores of rupees)	Box Office Collection
2	Release Time	Categorical with 4 levels	Releasing_Time_Festival Season Releasing_Time_Holiday Season Releasing_Time_Long Weekend Releasing_Time_Normal_Season
3	Genre	Categorical with 5 levels	Genre_Action (Action) Genre_Drama (Drama) Genre_Romance (Romance) Genre_Comedy (Comedy) Genre_Others (Other-G)
4	Movie Content	Categorical with 3 levels	Masala (Masala) Sequel (Sequel) Others (Other_C)
5	Director Category	Categorical with 3 levels	Director_A Director_B Director_O
6	Lead Actor Category	Categorical with 3 levels	Actor_A Actor_B Actor_O
7	Music Director Category	Categorical with 3 levels	Music_Dir_CAT A Music_Dir_CAT B Music_Dir_CAT C
8	Production House Category	Categorical with 3 levels	Prod_House_CAT A Prod_House_CAT B Prod_House_CAT C
9	Item Song	Binary variable	Item_Song (1 implies that the movie has an item song, 0 otherwise)
10	Budget	Numerical (in crores of rupees)	Budget
11	YouTube Views	Numerical	YouTube-V
12	YouTube Likes	Numerical	YouTube-L
13	YouTube Dislikes	Numerical	YouTube-D
14	Budget More than 35 crores	Categorical	Budget_35_Cr (1 if the budget is more than 35 crores 0 otherwise)

A simple linear regression model was developed between Box office collection and budget. SPSS output of the model is shown in Tables 2.2-2.3 and Figures 2.1-2.2.

Model 1

Y (Box Office Collection) = $\beta_0 + \beta_1$ x Budget

Table 2.2. Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.650[a]			72.02261

a. Predictors: (Constant), Budget

b. Dependent Variable: Box_Office_Collection

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8.354	8.535		-.979	.329
	Budget	2.175	.210	.650	10.381	.000

a. Dependent Variable: Box_Office_Collection

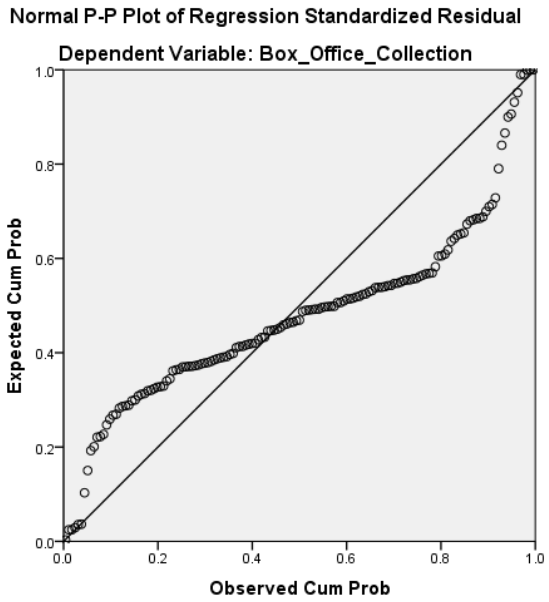


Figure 2.1. Normal P_P plot for Model 1

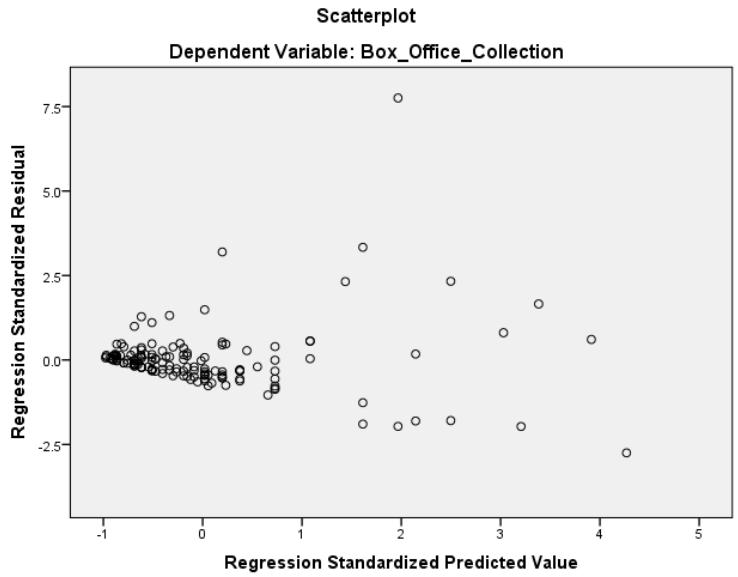


Figure 2.2. Residual plot for Model 1

Question 2.1 (2 points – only when all correct answers are identified)
Which of the following statements are correct (more than one may be correct)?
Tick (√) all right answers or highlight the correct statements with color.

1. The model explains 42.25% of variation in box office collection.
2. There are outliers in the model.
3. The residuals do not follow a normal distribution.
4. The model cannot be used since R-square is low.
5. Box office collection increases as the budget increases.

Question 2.2 (3 Points)
Mr Chellappa, CEO of Oho Productions (OP) claims that the regression model in Table 2.3 is incorrect since it has negative constant value. Comment whether Mr Chellappa is correct in his assessment about the model.

A second model is developed between ln(Box office collection) and movie release time:

Model 2

$$\ln(Y) = \beta_0 + \beta_1 \times \text{Release Time Festival Season} + \beta_2 \times \text{Release Time Long Weekend} + \beta_3 \times \text{Release Time Normal Season} + \varepsilon$$

The regression output for Model 2 is given in Table 2.4.

Table 2.4 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	2.685	.396		6.776	.000
	Releasing_Time_Festival_Season	.727	.568	.136	1.278	.203
	Releasing_Time Long_Weekend	1.247	.588	.221	2.122	.036
	Releasing_Time Normal_Season	.147	.431	.041	.340	.734

a. Dependent Variable: Ln(Box Office Collection)

Question 2.3 (3 points)
What is the average difference in the box office collection when a movie is released during a holiday season (Releasing_Time_holiday_season) versus movies released during normal season (Releasing_Time_Normal_Season)? Use a significance value of 5%.

Question 2.4 (4 Points)
Mr Chellappa of Oho productions claims that the movies released during long weekend (Releasing_Time_Long_Weekend) earn at least 5 crores more than the movies released during normal season (Releasing_Time_Normal_Season). Check whether this claim is true (use α = 0.05).

A stepwise regression model is developed between ln(Box Office Collection) and all the predictor variables listed in Table 2.1. The outputs are shown in Tables 2.5-2.6.

Table 2.5 Model Summary[g]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.709[a]	.503	.499	1.20651
2	.763[b]	.581	.576	1.11050
3	.787[c]	.620	.612	1.06210
4	.802[d]	.643	.633	1.03307
5	.810[e]			1.01749
6				

Table 2.6. Coefficients in the model (in the order in which it was added to the model)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Correlations		
		B	Std. Error	Beta		Zero-order (direct)	Partial	Part
Step 6	(Constant)	3.573	.249		14.346			
	Budget_35_Cr	1.523	.207	.443	7.342	.709	.525	.356
	Youtube_Views	1.1710^{-07}	.000	.242	4.426	.538	.348	.214
	Prod_House_CAT A	.562	.185	.165	3.033	.444	.247	.147
	Music_Dir_CAT C	-.645	.199	-.177	-3.245	-.483	-.263	-.157
	GenreComedy	.456	.197	.115	2.312	.006	.190	.112
	Director_CAT C	-.434	.203	-.123	-2.143	-.509	-.177	-.104

Question 2.5 (2 Points)

What is the variation in response variable, ln(Box office collection), explained by the model after adding all 6 variables?

Question 2.6 (2 Points)

Which factor has the maximum impact on the box office collection of a movie? What will be your recommendation to a production house based on the variable that has maximum impact on the box office collection?

Question 2.7 (2 Points)

Compare the regressions in Model 2 (Table 2.4) and Model 3 (Tables 2.5 and 2.6). None of the variables in Model 2 are statistically significant in Model 3. Can we conclude that the variables in Model 2 have no association relationship with Box Office Collection? Explain clearly.

Question 2.8 (2 Point)

Among the variables in Table 2.6, which variable is not useful for practical application of the model? Clearly state your reasons.

Question 3: (20 Points)

A data analytics start up works with political parties during elections. They have got access to voting patterns from various official sources. They are trying to understand how the percent of votes obtained by the winner is determined. As a first cut they are using the following data:

% VOTES – the percent of votes polled obtained by the winning candidate

MARGIN – the margin of victory measured in number of votes

Gender – 1 is for Men and 0 for women

College – 1 is for college educated winners and 0 for those who did not go to college.

They run the regression for all 543 elected MPs. The model output is provided below (with few missing information):

Table 3.1

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	
Observations	543

Table 3.2 ANOVA

	Df	SS	MS	F	Significance F
Regression					
Residual		17104.06			
Total	542	36481.89			

Table 3.3 Coefficients						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	38.59235	0.937225			36.75129	40.4334106
MARGIN	5.32E-05	2.18E-06			4.89E-05	5.7463E-05
Gender	1.551306	0.777806			0.023404	3.07920835
College	-1.47506	0.586995			-2.62814	-0.3219783

3.1 Fill up the Tables 3.1, 3.2 and 3.3 above (except the p values and the Significance F values). Clearly write all the steps. [10 points]

3.2 Assuming that 't' is significant for any value greater than 1.964 at 5%, Are the variables (margin, gender and college) significant? [2 points]

3.3 Assuming that the critical value of F is 2.621 at 5% significance, Is the overall regression significant? (2 points)

The analytics firm decides to dig a little deeper and looks at two outlying states, UP and AP, one of which has significantly lower assets per winner and the other significantly higher. Both the new variables are 0-1 variables. The values for some of the regressions are given below (Table 3.4).

Table 3.4 Regression Models with Corresponding R-Square

Regression Model	Independent Variables	R^2
1	MARGIN	
2	MARGIN, Gender	0.52567
3	MARGIN, Gender, College	0.531163
4	MARGIN, Gender, College, UP	0.56051
5	MARGIN, Gender, College, UP, AP	0.581339

3.4 What is the part correlation for College and % of votes in Regression model 3? [2 points]

3.5 Between regression 2 and 5, Is it justified to add the additional variables? (2 points)

Regression model 5 in Table 3.4 has a standard error of 5.333135, an overall F value of 149.1324 with significance of $4.4x10^{-99}$. The standard deviation for the dependent variable is 8.204253. The values of standard deviation for the dependent and independent variables are given below (Table 3.5).

Table 3.5

	Coefficients	Standard deviation
Intercept	38.56993	
MARGIN	5.58E-05	111365.7
Gender	1.498308	0.311494
College	-1.53774	0.412796
UP	-3.71439	0.354761
AP	5.715821	0.209766

3.6 Which variable has the greatest impact on voting %? (2 points)

Question 4 (20 Points)

The Indian life insurance market is covered by 24 life insurance companies with Life Insurance Corporation (LIC) as the only public sector company and rest are private with combination local and global companies. The insurance companies are focused in improving persistency for a life insurance policy where retaining the customer is an important priority. However usually less than half the policies sold survive beyond the second policy year. Low rates of policy persistency in the face of lower growth in coverage are not healthy for the industry.

Data on 2000 policies were collected from an insurance company and the data description is shown in Table 4.1.

Table 4.1 Description

S.No	Variable	Variable Type	Code in SPSS output
1	Renewal Outcome (Y)	Categorical	1 = Renewal Paid 0 = Renewal Not Paid
2	Plan Type	Categorical	1 = Savings Plan 0 = Protection Plan
3	Annual Premium	Numerical	Premium
4	Sum Assured	Numerical	Sum Assured
5	Renewal Frequency	Categorical with 4 categories	ANNUAL HALF YEARLY QUARTERLY MONTHLY
6	Payment Method	Binary Variable	1 = ECS 0 = Cheque
7	Policy Term	Numerical	Policy Term
8	Gender	Categorical	1 = Male 0 = Female
9	Customer Annual Earning	Numerical	Annual.Earning
10	OCCUPATION	Categorical	Agriculture Salaried Business House Wife Professional Others
11	Education	Categorical	1. Below SSC 2. SSC 3. HSC 4. Graduate 5. Others
12	Marital Status	Categorical	1. Single 2. Married 3. Divorced 4. Widowed

13	Premium to Income ratio	Numerical	Premium to Income

A logistic regression model is developed between Renewal Payment and Annual Premium. That is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \times \text{Premium} \qquad \text{(Model 1)}$$

The model outputs are shown in Tables 4.2-4.5 and Figure 4.1.

Table 4.2. Iteration history

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients	
			Constant	Premium
Step 1	1	2656.933	-.320	.000
	2	2630.622	-.589	.000
	3	2629.321	-.658	.000
	4	2629.318	-.661	.000
	5	2629.318	-.661	.000

a. Method: Forward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 2768.537

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Table 4.3. Omnibus Test

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.219	1	.000
	Block	139.219	1	.000
	Model	139.219	1	.000

Table 4.4 Classification Table

Classification Table[a]

Observed			Predicted		
			Renewal Outcome		Percentage Correct
			0	1	
Step 1	Renewal Outcome	0	751	204	78.6
		1	353	692	66.2
	Overall Percentage				72.2

a. The cut value is .500

Table 4.5 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Premium	.000	.000	95.580	1	.000	1.000
	Constant	-.661	.085	60.060	1	.000	.516

a. Variable(s) entered on step 1: Premium.

(4.1) Calculate the value of – 2LL0. What is the interpretation of -2LL0? After adding the variable "Premium", what is the reduction in -2LL? (2 Points)

(4.2) Calculate the precision for the model when the classification cut-off probability is 0.50. (2 point)

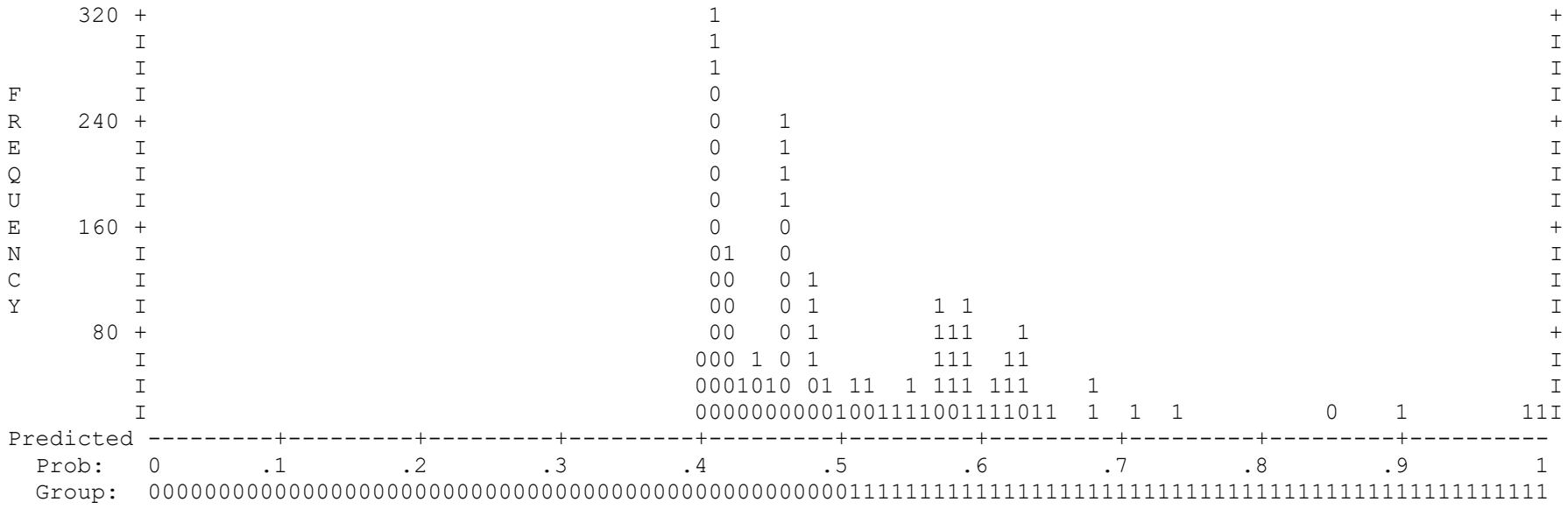
(4.3) Is the variable "Premium" statistically significant at $\alpha = 0.01$? Observe that exp (β) = 1 in Table 4.5. (1 point)

(4.4) The management decided to call the customers through their call centre whose probability of premium payment is less than 0.5. How many calls will be made? (1 point)

(4.5) Calculate approximate sensitivity, specificity, precision and F-score for cut-off probability of 0.8. (3 points)

Step number: 1

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for 1
The Cut Value is .50
Symbols: 0 - 0
1 - 1
Each Symbol Represents 20 Cases.

Figure 4.1. Classification plot for model 1

A second model is developed using "Occupation" as the predictor and the regression model is shown in Table 4.6.

Table 4.6 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Salaried	1.927	.102	358.118	1	.000	6.866
	Constant	-.738	.064	131.361	1	.000	.478
Step 2[b]	Salaried	2.008	.103	376.956	1	.000	7.451
	Others	1.603	.309	26.890	1	.000	4.968
	Constant	-.820	.067	149.940	1	.000	.440
Step 3[c]	Salaried	2.046	.104	387.018	1	.000	7.740
	House Wife	22.061	11602.711	.000	1	.998	3811265942.657
	Others	1.641	.309	28.148	1	.000	5.161
	Constant	-.858	.068	159.882	1	.000	.424

a. Variable(s) entered on step 1: Salaried.

b. Variable(s) entered on step 2: Others.

c. Variable(s) entered on step 3: HouseWife.

4.6 Calculate the probability that a house wife will pay the renewal. (1 Point)

4.7 Customers with which occupations have low probability of renewal payment? (1 point)

Table 4.7 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 12[k]	Business	-.538	.126	18.239	1	.000	.584
	Professional	-2.425	.794	9.335	1	.002	.088
	Premium	.000	.000	9.877	1	.002	1.000
	ANNUAL	-3.904	.246	251.341	1	.000	.020
	HALFYEARLY	-3.610	.277	169.295	1	.000	.027
	QUARTERLY	-2.396	.439	29.796	1	.000	.091
	Policy Term	-.026	.004	36.957	1	.000	.975
	Graduate	.480	.137	12.233	1	.000	1.615
	Others	.782	.191	16.717	1	.000	2.186
	Married	.437	.163	7.215	1	.007	1.548
	Constant	3.105	.274	128.827	1	.000	22.313

Table 4.7 shows the model developed using stepwise regression.

Use Table 4.7 to answer questions 4.8 and 4.9.

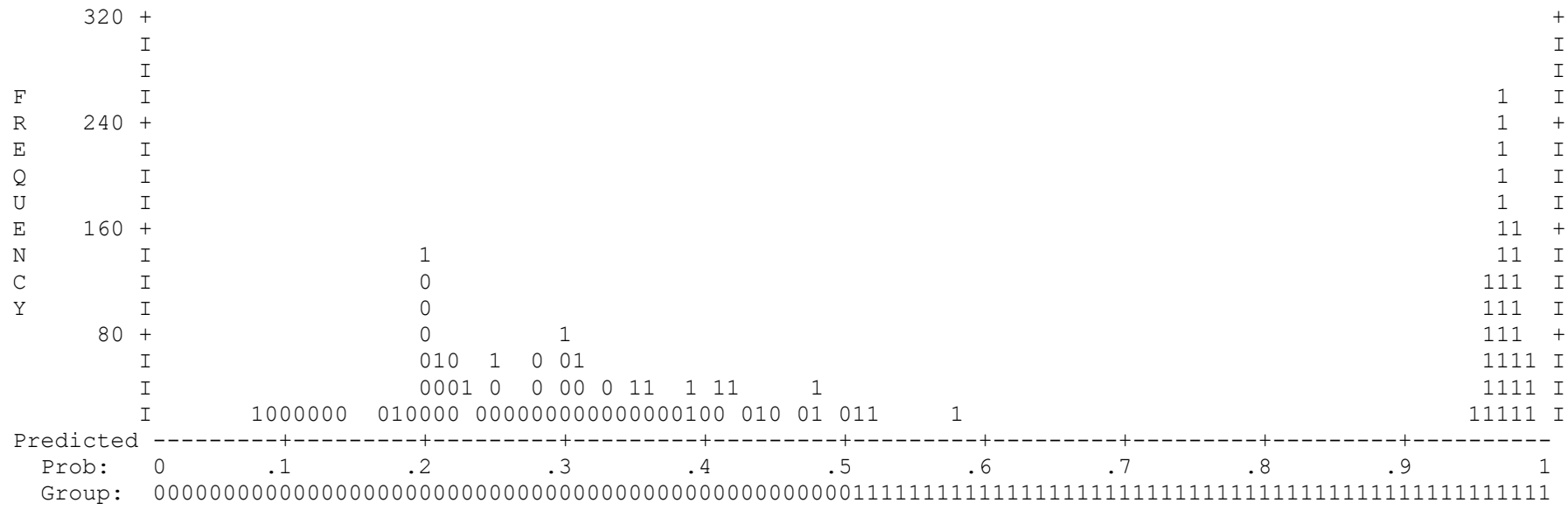
4.8) Calculate the probability of renewal payment of a customer with following features: 1. Agriculturist, 2. Monthly premium, 3. with a policy term of 60, 4. SSC education and 5. Marital status single (ignore premium amount) (2 points).

4.9) What will be your recommendation to the insurance company based on the model in Table 7. Discuss possible deployment strategies (2 points).

4.10) Use the classification plot in Figure 4.2 and calculate approximate gain and lift for the first 5 deciles. How do you interpret lift value, what is your recommendation based on life value (5 points)

Step number: 11

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for 1
The Cut Value is .50
Symbols: 0 - 0
1 - 1
Each Symbol Represents 20 Cases.

Figure 4.2. Classification plot for full model

Question 5 (20 Points)

Go through the case, "Oakland A" and the spreadsheet supplement (Ref: Moodle/Cases and Materials/Module 3). Does mark Nobel increase attendance? If so, how much is the increase worth for Oakland? Support your decision through an appropriate regression model.

Question 6 (20 Points)

Read the case, "Predicting Net Promoter Score (NPS) to Improve Patient Experience at Manipal Hospitals". Answer the following questions:

1. What is the business problem in this case and how is this business problem converted into an analytics problem? (3points)
2. How can we estimate sensitivity and specificity of three-class problem? (4points)
3. Develop a binary logistic regression model by combing detractors and passive into one class, what will be your advice to Manipal hospital based on the model developed? (5 points)
4. Develop an ordinal logistic regression on a three-class problem (Detractor/Passive/ Promotor)? Compare the ordinal regression model with logistic regression model developed in question 3. (6 points)
5. What will be your recommendation to Manipal hospital? (2 points)