Assigment – Rohit Garg

1. Assignment 1

1.1. Objective

The credit risk director would like to know which loans are going to default. There are two datasets – applicant and loans.

1.2. Data

The model is developed on 1,000 observations and 18 independent variables. There are 700 (70%) non-defaults and 30 (30%) defaults. Based on information value (IV) 6 independent variables are shortlisted. IV indicates the explanatory power of the independent variables. **Please refer appendix D for details on variable shortlisting.**

- **Categorical variables –** Based on the default rate the categories are combined together. The chart below shows the default rate for the combined categories.
- **Numerical importance –** Decision trees are used to bin the numerical variables (max depth of the tree is 3 and min samples leaf is 50). Based on the default rate the bins are combined together. The chart below shows the default rate for the combined bins.

loan_amount	value
0 to 13.7 Lakhs	0.32
13.7 Lakhs to 34.5 Lakhs	0.24
34.5 Lakhs to 39.1 Lakhs	0.08
39.1 Lakhs to 50.0 Lakhs	0.48
50.0 Lakhs to 78.2 Lakhs	0.32
78.2 Lakhs or more	0.54

Loan_history	Value
all loans at this bank paid back duly	0.57
critical/pending loans at other banks	0.18
delay in paying off loans in the past	0.30
existing loans paid back duly till now	0.30
no loans taken/all loans paid back duly	0.63

months	value
1 to 8 months	0.10
9 to 11 months	0.20
12 to 15 months	0.24
16 to 33 months	0.32
34 to 42 months	0.42
43 or more months	0.56

Purpose	Value
business	0.36
career development	0.12
domestic appliances	0.36
education	0.45
electronic equipment	0.21
FF&E	0.36
new vehicle	0.36
repair costs	0.36
used vehicle	0.18
missing	0.45

age	value
23 or less	0.39
24 to 25	0.45
26 to 27	0.27
28 to 34	0.33
35 to 36	0.15
37 or more	0.27

1.3. Insights

We have developed a logistic regression model to identify the key variables that help to identify if an applicant will default or will not default. The model has an AUROC of 0.774 and cost of 487.

	coef	std err	z	P>|z|
const	-7.812	0.634	-12.325	0.000
loan_amount	3.684	0.736	5.008	0.000
months	3.734	0.743	5.028	0.000
history	4.286	0.693	6.182	0.000
age	5.343	1.093	4.890	0.000
purpose	5.442	0.970	5.613	0.000

1.4. User Interface

- **Inputs –**Loan Amount, Loan Duration (Months), Loan History, Age of the applicant and Purpose
- **Output –** The Probability of default (PD) and Decision. If PD > 0.2 then BAD else GOOD. The cut-off is decided based on the cost matrix
- **Please refer appendix E for details on the user interface.**

2. Assignment 2

2.1. Objective

The marketing director would like to know which holiday brings in the most money so the team can adjust the marketing dollars.

2.2. Data

The e-commerce data is read and rolled up at year-week and country level. It is observed that the data is only for 2 years – 2010 and 2011. There are two dependent variables:

- **Quantity –** it is the sum of quantity for a given year-week and country
- **Total price –** it is the sum of quantity x unit price for a given year-week and country

The US holiday dates data is read. The date is converted to year-week. Following two data treatments steps are applied:

- **Holiday week –** Spread of holidays
    - If the holiday falls on Monday or Tuesday (start of week) then the holiday is spread across two weeks the previous week and the current week.
    - If the holiday falls on Wednesday or Thursday or Friday then the holiday is only considered for current week.
    - If the holiday falls on Saturday or Sunday (end of week) then the holiday is spread across two weeks the current week and the next week.
- **Combined holidays –** It is observed that Christmas Day and Christmas Eve can be combined, Western Easter and Eastern Easter can be combined, Thanksgiving Day and Thanksgiving Eve can be combined, Labour Day Weekend and Labour Day can be combined.

2.3. Insights

For the entire portfolio, the data is rolled up at year-week level and models are run. Based in the p-value (level of significance) the holidays are shortlisted. There are two models:

- **Quantity –** Christmas, Easter and Memorial
- **Total price –** Christmas and Easter

Hence, for the entire portfolio there are 3 important holidays – Christmas, Easter and Memorial. Just before the start of these holidays there is significant increase in the quantity and total price. Hence the marketing director should increase the marketing spend weeks before these holidays. **Please refer appendix A for the charts**

For individual country, the data is rolled up at rolled up at year-week and country level. Based in the p-value (level of significance) the holidays are shortlisted. Based on the two models the following countries are shortlisted:

- **France, Ireland and Spain –** weeks before New Year and Labor Day the marketing director should increase the marketing spend
- **Channel Islands and Japan –** weeks before Valentine the marketing director should increase the marketing spend
- **Australia and Norway –** weeks before Juneteenth and Columbus Day the marketing director should increase the marketing spend
- **Belgium, Cyprus, Germany, Portugal and United Kingdom –** weeks before Columbus Day the marketing director should increase the marketing spend
- **Italy –** weeks before Thanksgiving the marketing director should increase the marketing spend

Please refer appendix B for details on the matrix between country and holiday.

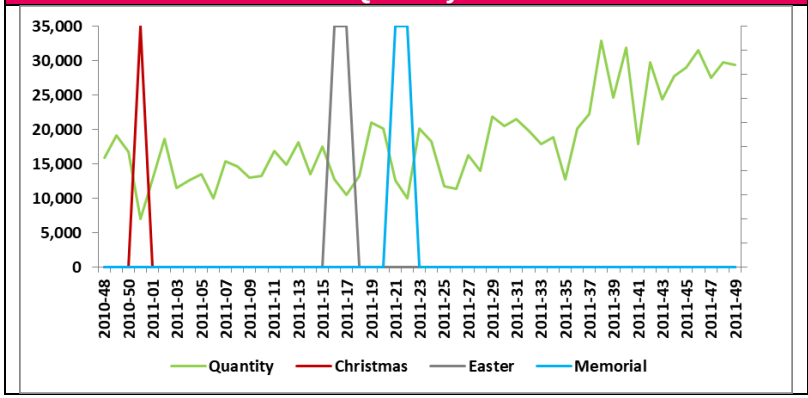
2.4. User Interface

Please use the drop down menu to select the country and holiday. The user can select up to 2 countries and up to 3 holidays. Based on the selection the chart (Total Price) is updated. **Please refer appendix C for details on the user interface.**

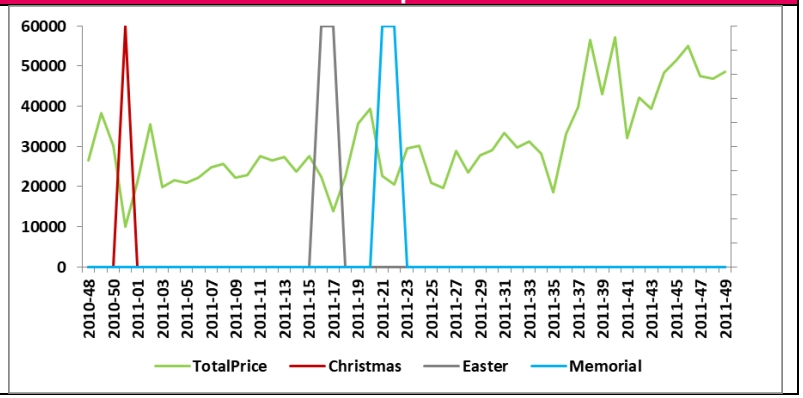
3. Appendix

3.1. A – Charts

Quantity



Total price



3.2. B – Combinations

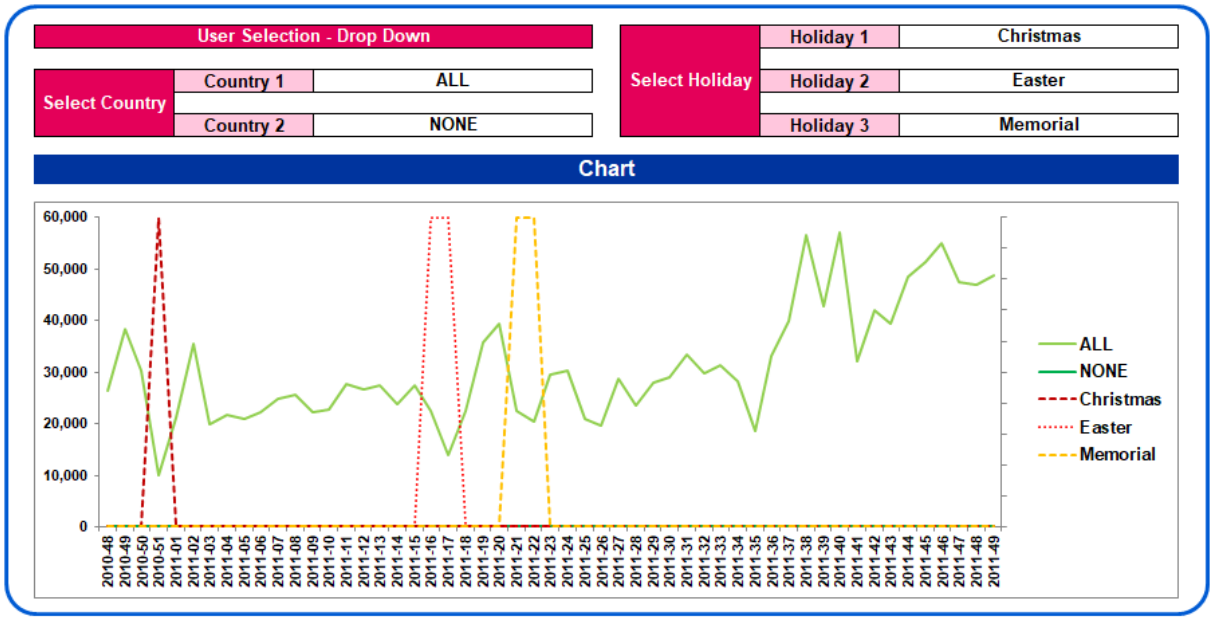
Quantity

	France, Ireland and Spain	Japan	Australia	Belgium, Germany and Portugal	Italy	Others
Christmas						
New Year	Action					
Valentine		Action				
Easter						
Memorial						
Juneteenth			Action			
4thJuly						
Labor Day	Action					
Columbus Day				Action		
Veterans Day						
Thanks giving					Action	

Total price

	France, Ireland and Spain	Japan and Channel Islands	Australia and Norway	Belgium, Cyprus, Germany, Portugal, United Kingdom	Italy	Others
Christmas						
New Year	Action					
Valentine		Action				
Easter						
Memorial						
Juneteenth			Action			
4thJuly						
Labor Day	Action					
Columbus Day			Action	Action		
Veterans Day						
Thanks giving					Action	

3.3. C – User Interface



3.4. D – Variable Funnel

Raw Variables (18 variables)	Shortlist based on IV (6 variables)	Shortlist based on Logistic Reg (5 variables)
age	age	age
balance	history	history
coapplicant	loan_amount	loan_amount
dependents	months	months
emi_pct	purpose	purpose
emi_plan	**property**	
employ		
employ_year		
guarantor		
history		
house		
house_years		
loan_amount		
months		
property		
purpose		
sex		
status		

Additional variables (good to have):

- External rating (USA – FICO Score, India – CIBIL Score, …)
- Interest rates
- Location – Metro city, Non-metro city, rural
- Occupation and Income
- Collateral

3.5. E – User Interface

Outputs	
Probability of Default	0.738
Decision	**Default**
Inputs	
Loan Amount	0 to 13.7 Lakhs
Loan Duration (Months)	1 to 8 months
Loan History	all loans at this bank paid back duly
Age of the applicant	24 to 25
Purpose	missing