Data Story Telling for Capstone project – Dipanjan

#### DataSet Information:

This research is aimed at the case of customers default payments in Taiwan in 2005. The Credit Card Default Data on the UCI Machine Learning Repository can be found directly below:

<http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable.

This study reviewed the literature and used the following 23 variables as explanatory variables:  
**LIMIT\_BAL**: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.  
**SEX** : Gender (1 = male; 2 = female).  
**EDUCATION**: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).  
**MARRIAGE** : Marital status (1 = married; 2 = single; 3 = others).  
**AGE** : Age (year).

*(X6 - X11)*  
**PAY\_0 - PAY\_6**: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

*X12-X17*  
**BILL\_AMT1 - BILL\_AMT6**: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

*X18-X23*  
**PAY\_AMT1 - PAY\_AMT6**: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Data Story Telling:

The main goal of my investigation is to figure out, which variables among the 23 have higher impact in predicting probability of default next month. I took a series of steps and explored data to figure out the impacts:

1. Do age of the credit card holder have anything to do with default:
   1. The 40-49 age group was found to have higher percentage of defaulters that other age groups.
   2. The 60-69 age group was also found to be a bit vulnerable
2. Next I went on to check the last bill payment of the credit card holder. Does this amount say anything?
   1. The proportion of default is interestingly higher when the last payment amount is low.
   2. The default percentage keeps on going down as the last payment amount increases.
   3. The default percentage is higher if the last payment amount is below 10,000
3. Education impacts default:
   1. Those with just high school education tends to default more than those with graduate school or university education
4. Does marital status have an impact:
   1. My finding is that married people have slightly higher tendency to default than single folks.
   2. The marital status of ‘others’ have some interesting behavior of slightly higher rate of default
5. What about gender?
   1. Clearly, the males tend to default more than the females
6. I was currious to know whether credit has any relationship with default
   1. The finding is: card holders with lower credit credit limits tend to default more
   2. The credit limit ranges that showed more are ‘10000 – 50000’ and 50000- 100000’
7. Is a large number of missed payments a precursor to default?
   1. It is evident that if there are two or more missed payment, the chance of default increases significantly
   2. But, there is a strange observation. I have observed that there are defaults when there has been zero missed payments.
8. I also tried to find out correlations between all variables
   1. As expected, the missed payments and credit balance tend to indicate to some extent, that a default is coming.