Supevisor Meeting #2

MINUTES

JANUARY 29, 2015

SIS MR 4.1

MEETING CALLED BY	Nicholas Lee, Goh Jian Hao
TYPE OF MEETING	Supervisor
NOTE TAKER	Goh Jian Hao
ATTENDEES	Prof. Kam, Nicholas Lee, Goh Jian Hao

Topics discussed

DISCUSSION	Asked about the steps in EDA
Team had little experience in conducting EDA on a big dataset.	
CONCLUSIONS	Was informed of the usual procedures starting from univariate analysis (finding out the distributions, outliers, one-class distributions, patterns) for categorical variables and bivariate analysis for the predictors and discovering the extent of multicoliearity problem, which states the correlation between 2 or more predictors. Learned that for analysis, variables should follow a normally distributed pattern and if only its records are beyond the CLT of 30 records. Team understood that predictor variables should be minimally correlated with other predictors in order to produce a fair environment for predictive analysis. Subsequently, was apprised of data sampling and model construction steps. Validated with Supervisor that the Receiver operating characteristic curve will be a test for predictive power (false positives vs true positives)

DISCUSSION	Apprised Supervisor of the 2.23% of total records have missing values for races. Likewise, 60% for payer_code. Gave recommendations to remove them or to imput values based on probabilities.
CONCLUSIONS	Supervisor advised that imputation may introduce variation and errors and removal may cause relationships to be diminished, which may be crucial for analysis. Therefore, was advised to conduct 2-way analysis.

SPECIAL NOTES	Meeting ended in 45 minutes and supervisor and team are apprised of the above mentioned notes.