Modeling Summary

Data Preparation:

- Data clean up by fixing erroneous names, modifying variable formats and types, and imputing missing values. Check target leaks (not found).
- Binning most numeric variables based on weight of evidence: log(% target /(1-%target)). Created constant spline to capture the non-linear relations between the features and the target.
- Created dummy variables on all binning intervals.
- Use Random Forest to rank variable importance on the training data to select final 300 features.
- Plot some charts to show the correlations, target % by features. The charts demonstrate different pattern relationships between selected features and the target.

Model Development:

Two modeling methods are used in this exercise. One is the Gradient Boosting and another is the Stack model. Stack model is a combination of logistic regression, Naïve Bayes and GBM

Model 1: Gradient Boosting Machines (GBM) is a machine learning algorithm that produces a prediction model in the form of an ensemble of weak prediction models, decision trees. GBM can handle complicated relationships between the features and the target. The performance of GBM model is generally better than models that are based on statistical distribution assumption such as logistic regression. GBM has optimal performance but requires hyper parameter tuning and the interpretation could be challenging. It is also easier to cause overfitting.

Model 2: Stack model is a combination of logistic regression, Naïve Bayes and GBM. Logistic regression has been used in many industries to handle classification tasks. It is a good benchmark comparing to other modeling methods. The relationship between event odds ratio and the coefficients is straightforward. No hyper parameter tuning is necessary. However the model assumes that the predictors and the logit transformation of target variable is linear. Naïve Bayes does not need parameter tuning and is quick to produce outcomes. I used these two models to generate additional features. These additional features are the predictions from both models. For the final Stack model, I used GBM with these additional features.

Comparison:

Both models achieved high AUC values under ROC curve. In order to compare the performance of models, I randomly split the train data into modeling (67%) and validation (33%). Model 1 has validation AUC of 0.976 and Model 2 has validation AUC 0.973. The cumulative gain charts from logistic regression and Naïve Bayes illustrated the percentage of targets that are captured by prediction deciles.

Modeling Summary

The table and the charts below provide the feature importance rank and sample relationships of feature vs target. The cumulative gain chart from logistic regression shows that the top decile from the model score can catch at least 40% of the target.

Variable Name	Importance Rank
x97	0.021533
x75	0.021178
x58	0.018955
x37	0.01785
x41_float	0.015895
x99	0.013924
x1	0.013034
x53	0.01301
x63	0.012995
x66	0.012566
x83	0.012454
x21	0.012292
x44	0.012061

Modeling Summary