PERSONAL LOAN DEFAULT RISK PROJECT

Team member: Yen-Hsiu Chang ,Yule Jin, Zhen Li

This project is aimed at predicting the chance of delinquency over 90 days of the borrower on the loan in two years based on applicants' financial status and past credit activities.

DATA INTRODUCTION

- Data Sources
  - Kaggle.com – "Give Me Some Credit"
  - Observations: 150,000
  - Train data: 100,000, evaluation data: 50,000
  - Features: 10
- Data Attributes
  - Response - Dlqin2yrs (1,0)
  - Age, income, payment past due in the past, etc.
- Data preparation
  - Missing Values: K-NN

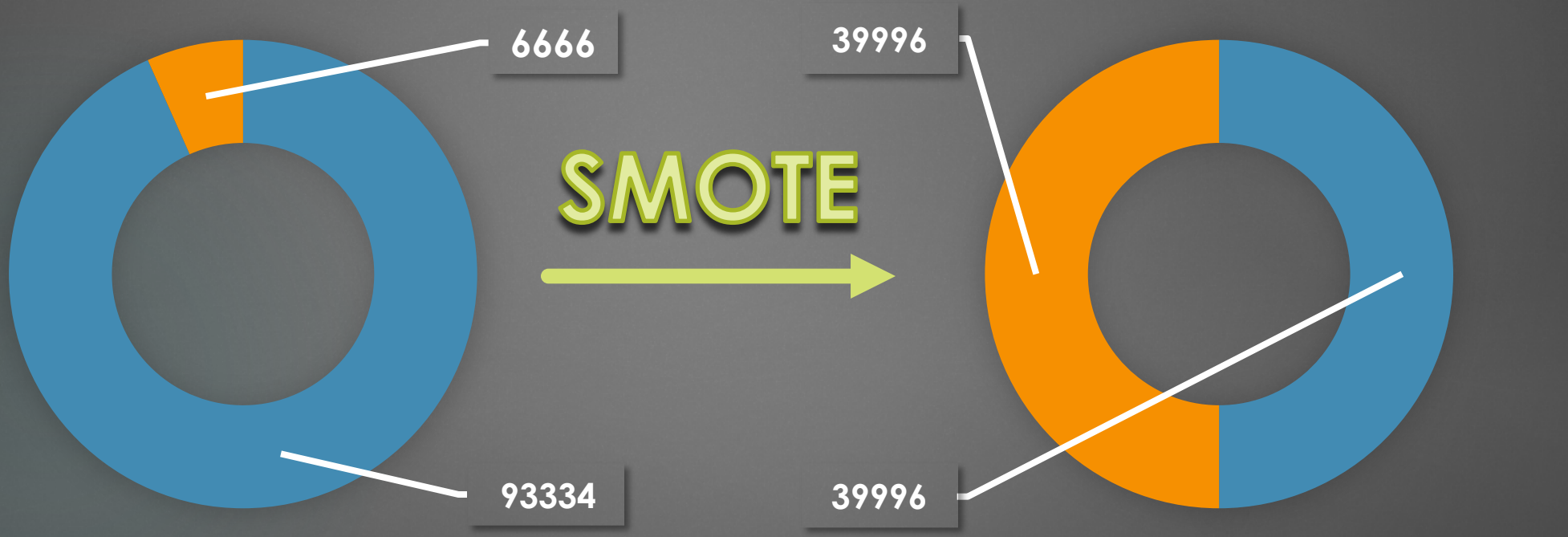
IMBALANCED DATASET

SOLUTION

SMOTE : Synthetic Minority Over-sampling Technique

PERFORMANCE MEASURE

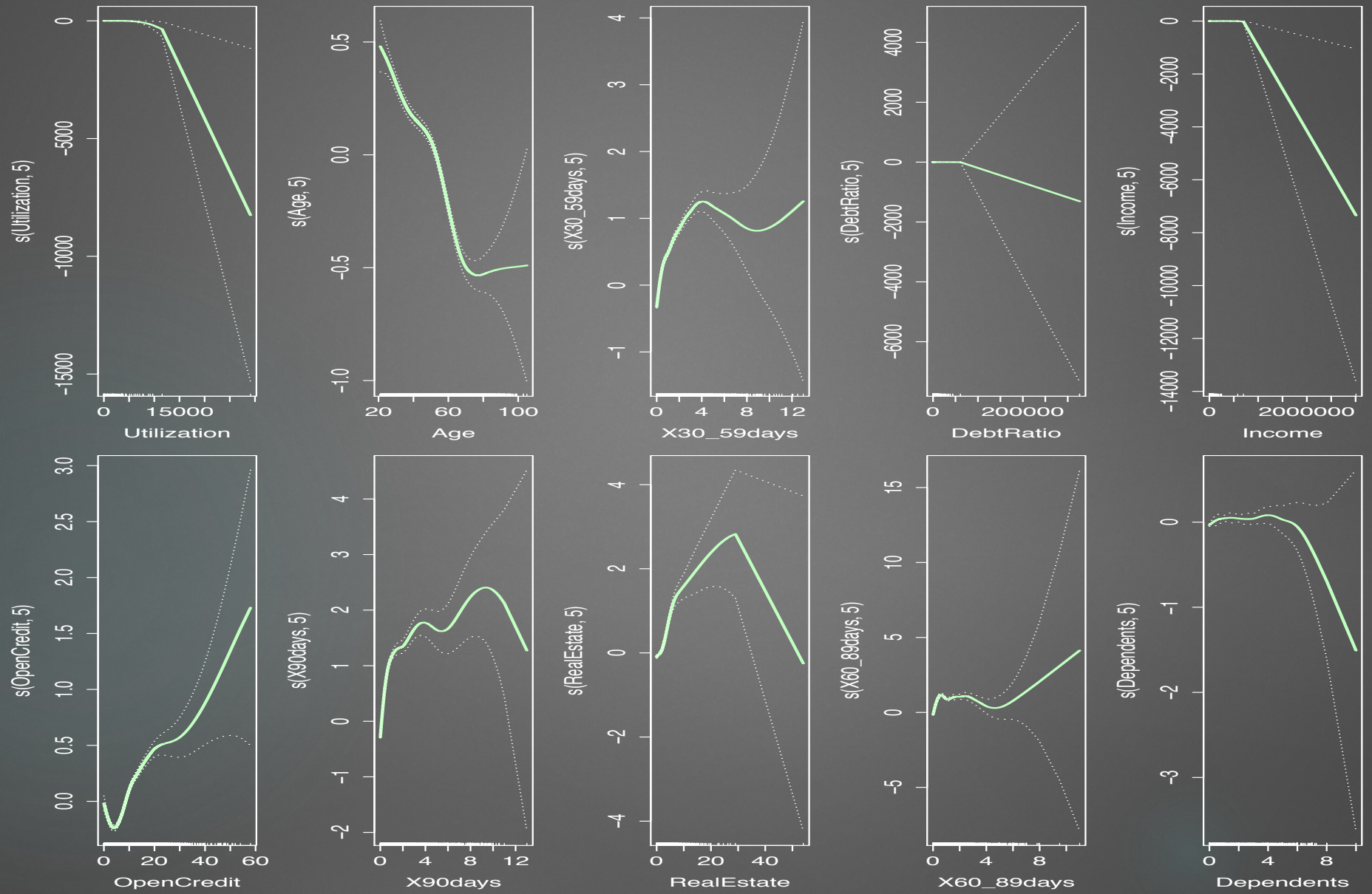
AUC: Area Under Receiver Operating Characteristic Curve



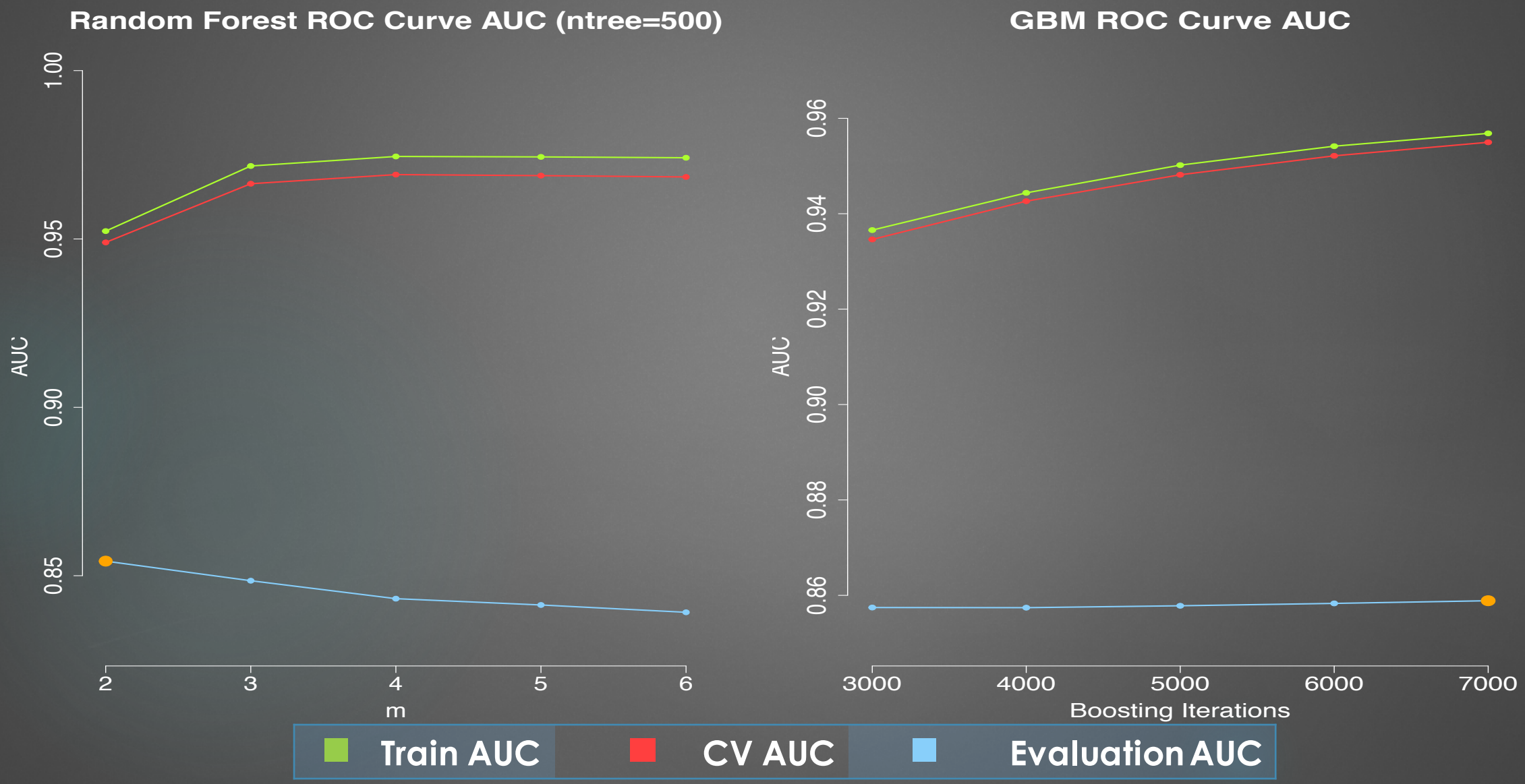
MODEL TRAINING

NAÏVE BAYES	*Kernel density estimate is used for density estimation*
LOGISTIC REGRESSION	
GAM	*Fitting smoothing splines (df = 5) for all features*
RANDOM FOREST	*Tuning m over (2 to 6) and number of trees (250 and 500)*
GBM	*Tuning number of trees from 3000 to 7000*

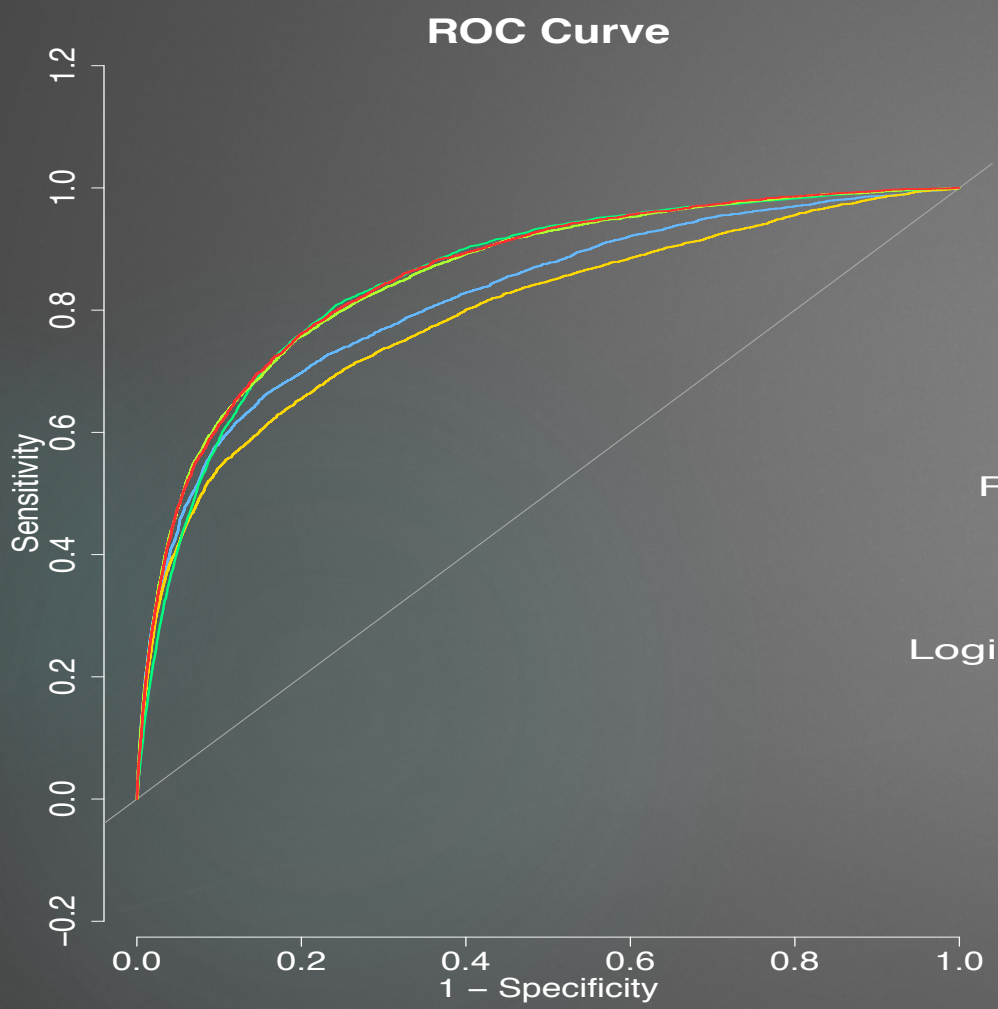
GAM FITTING RESULT

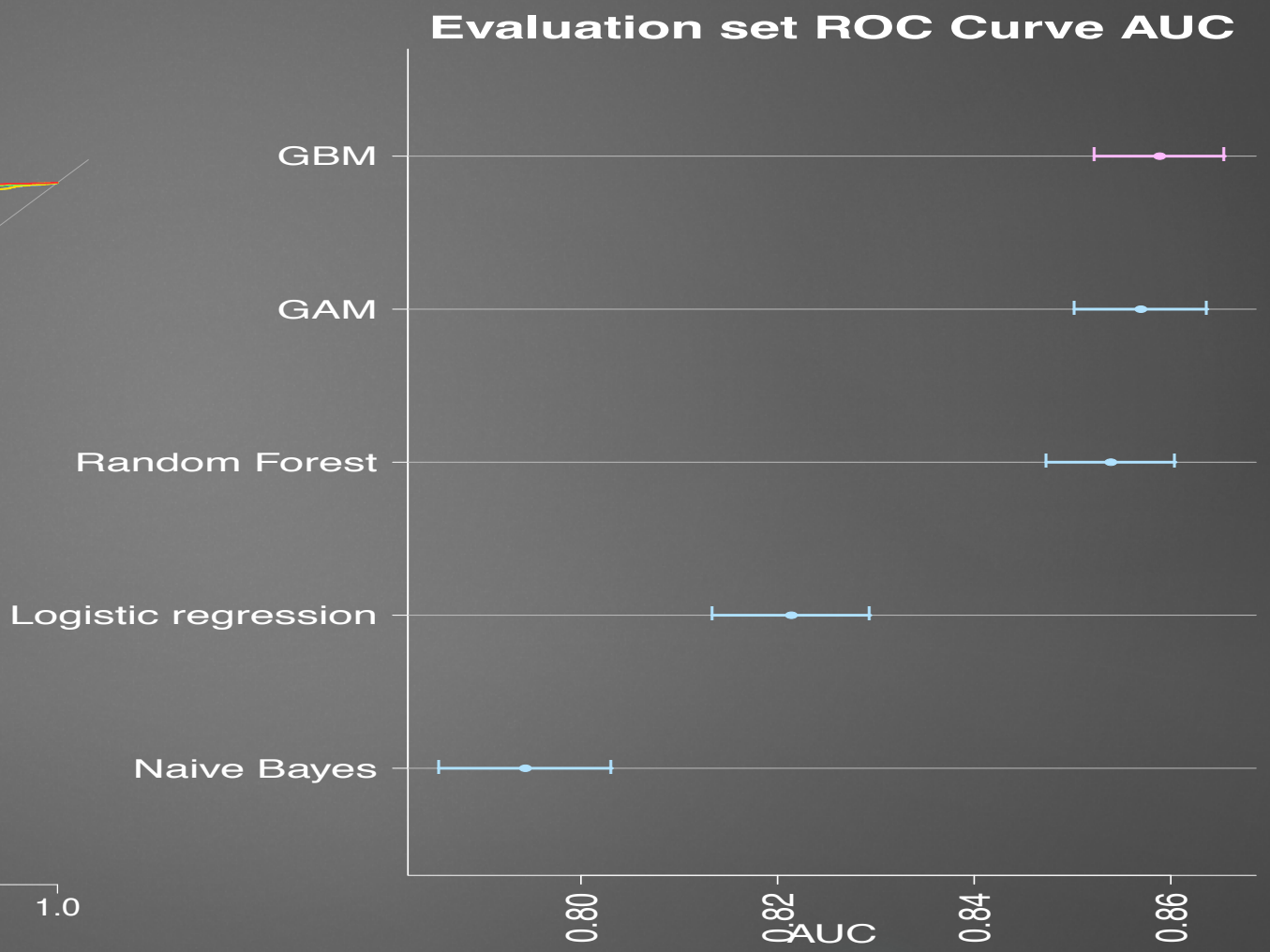


RANDOM FOREST & GBM OUTCOME



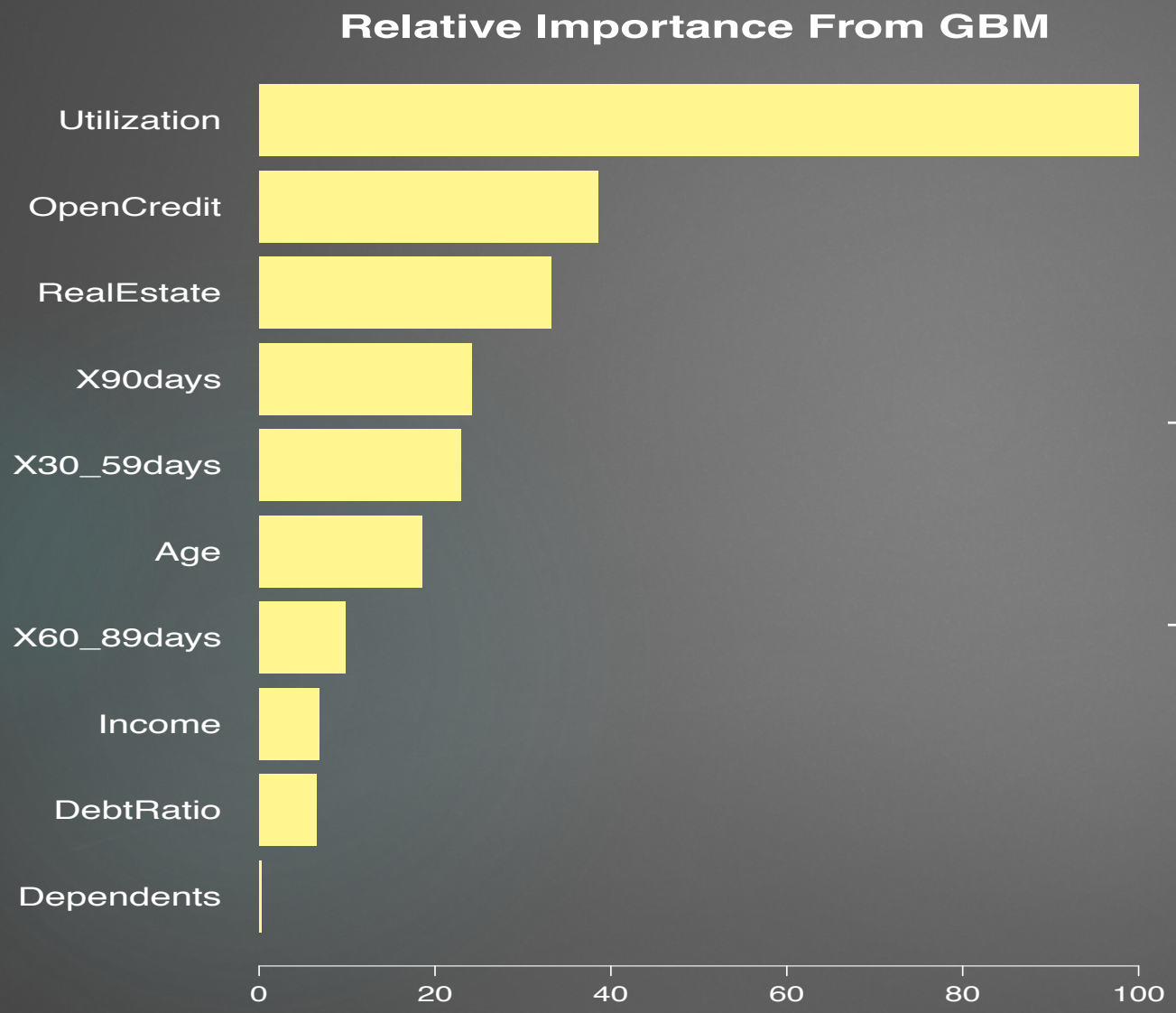
Summary





GBM GAM RANDOM FOREST LOGISTIC NAÏVE BAYES

Summary



Utilization	Total balance on credit cards and personal lines of credit except real estate and no installment debt divided by the sum of credit limits
OpenCredit	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)
RealEstate	Number of mortgage and real estate loans including home equity lines of credit