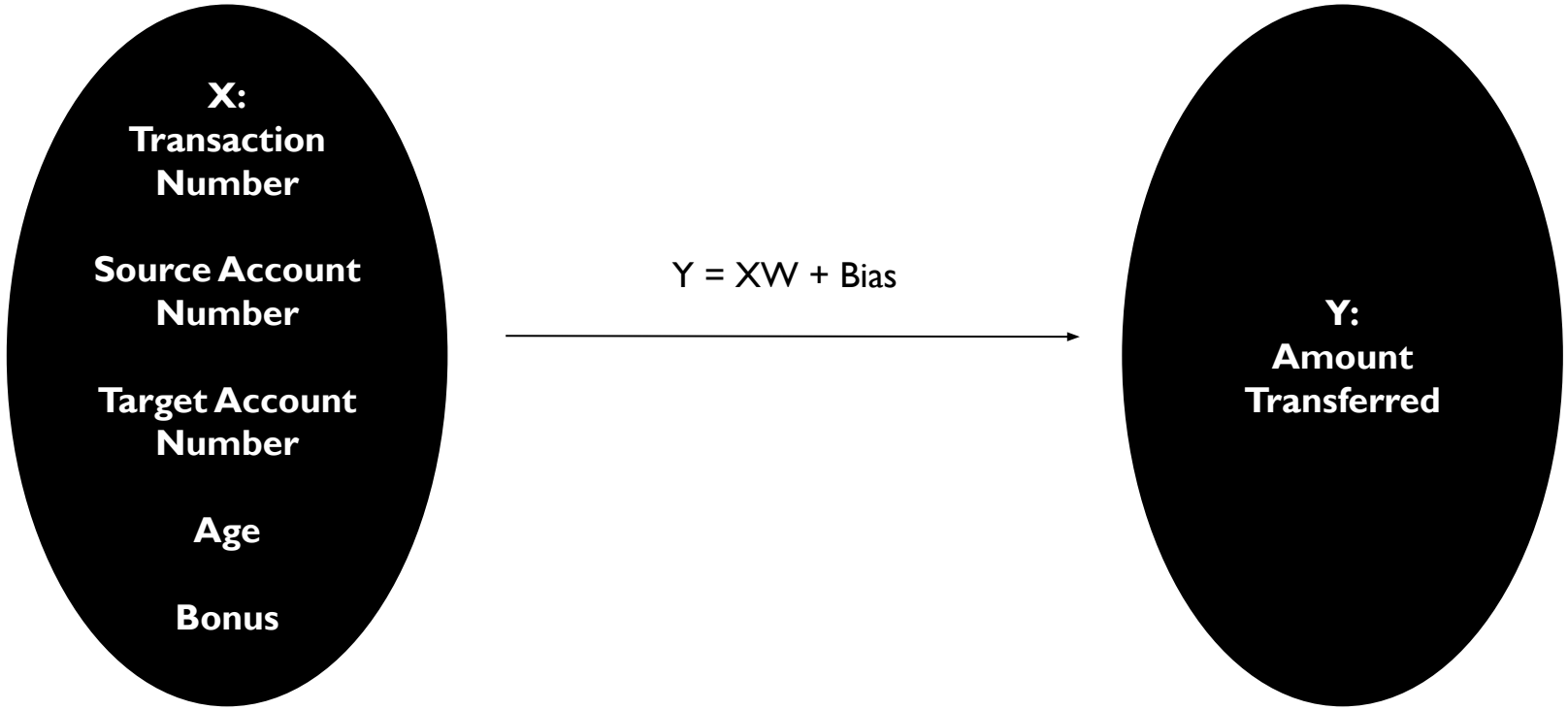
BDAD PROJECT

Team dataMiner

DATA GENERATION

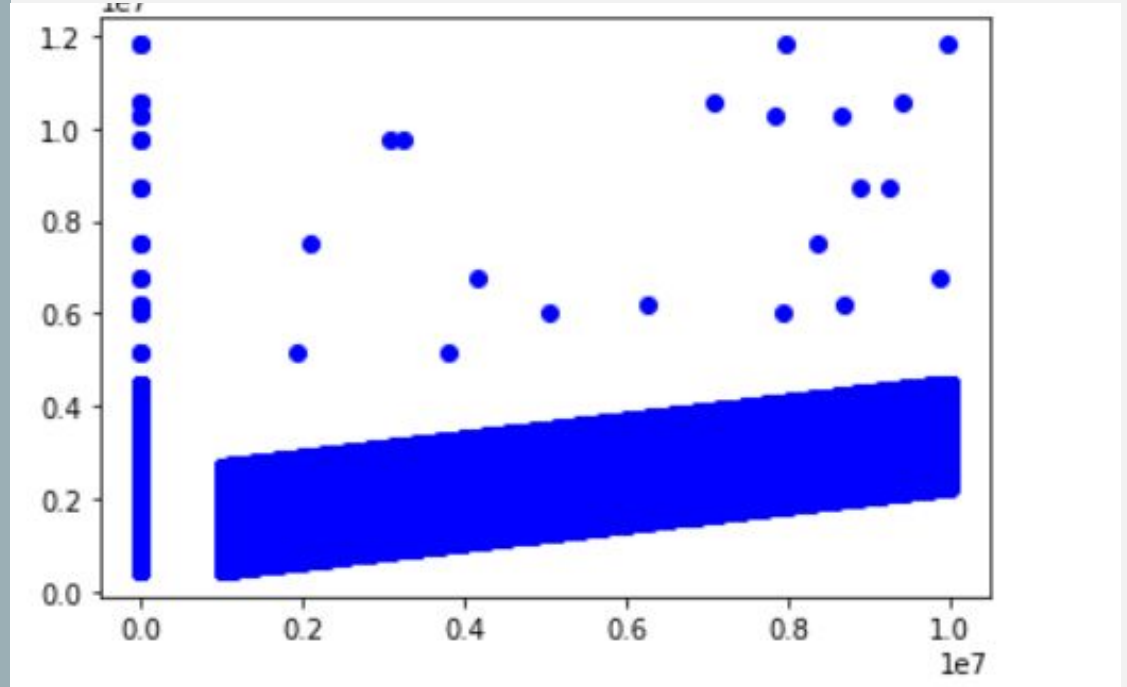


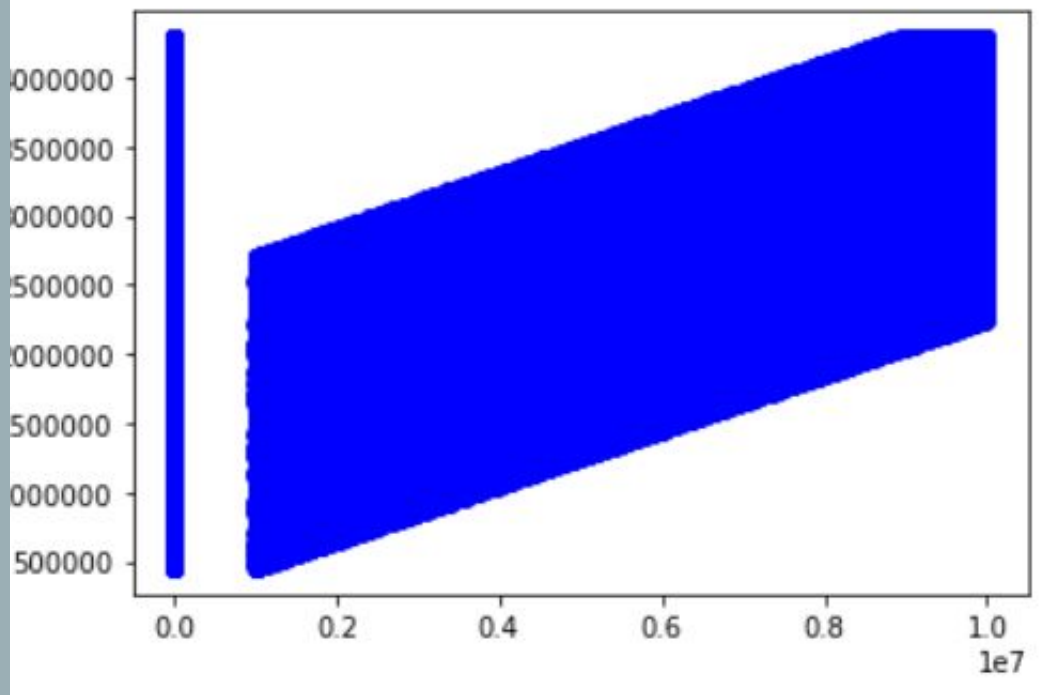
DEALING WITH NAN

- Two Strategies
  - Educated Guess: replace NaN with mean of value's column (chosen strategy)
  - Delete all NaN values

HANDLING OUTLIERS

- Set Upper limit threshold
- Quantify outlier strength
- Upper limit: (mean+stddev) * outlier strength





LR WITHOUT PCA

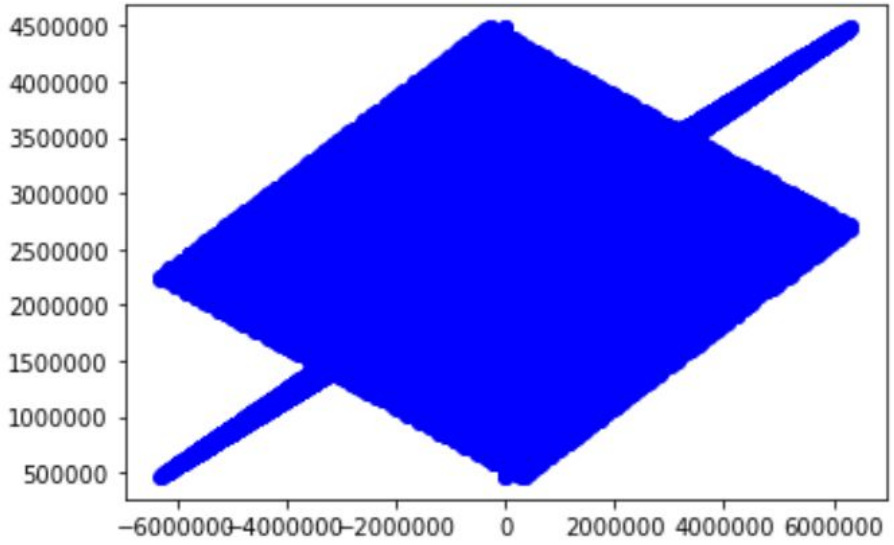
- Ran Linear Regression without PCA
- Evaluated with Normalized RMSE value of: 0.0068704268592233 02

PCA

- Ran PCA with two analyses:
  - Linear Regression (LR)
  - Gradient Boosted Tree (GBT)

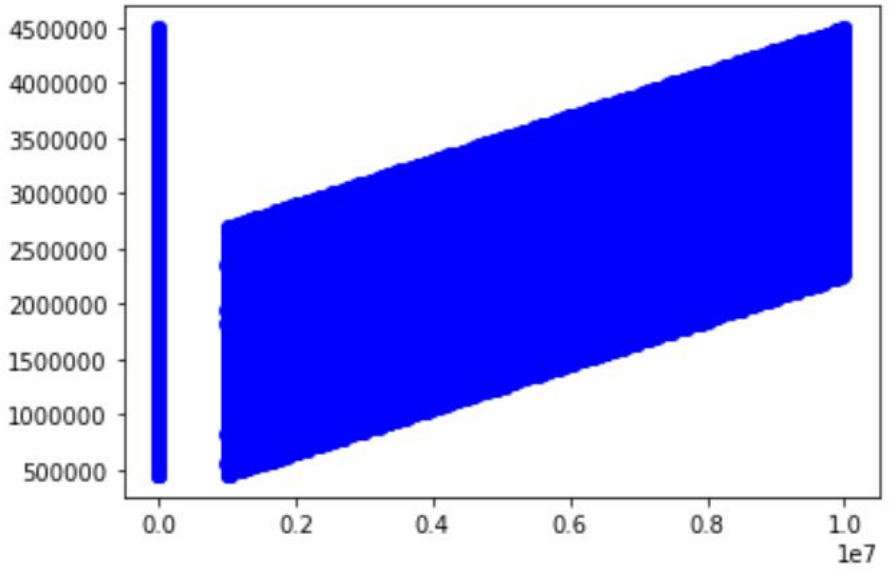
GBT PLOT

GBT Prediction on Test Set



LINEAR REGRESSION PLOT

LR Prediction on Test Set



RESULTS

- Normalized RSME for LR with PCA: 0.005543151988095 69
- Normalized RSME for GBT with PCA: 0.025395040320755412