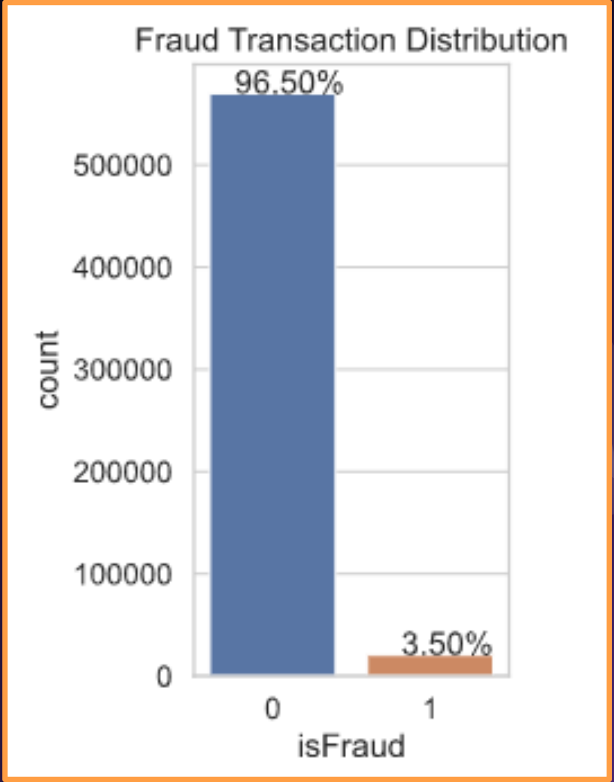
Calculating Fraud Risk in Credit Card Data

Rhys Carter
Metis, Spring 2021

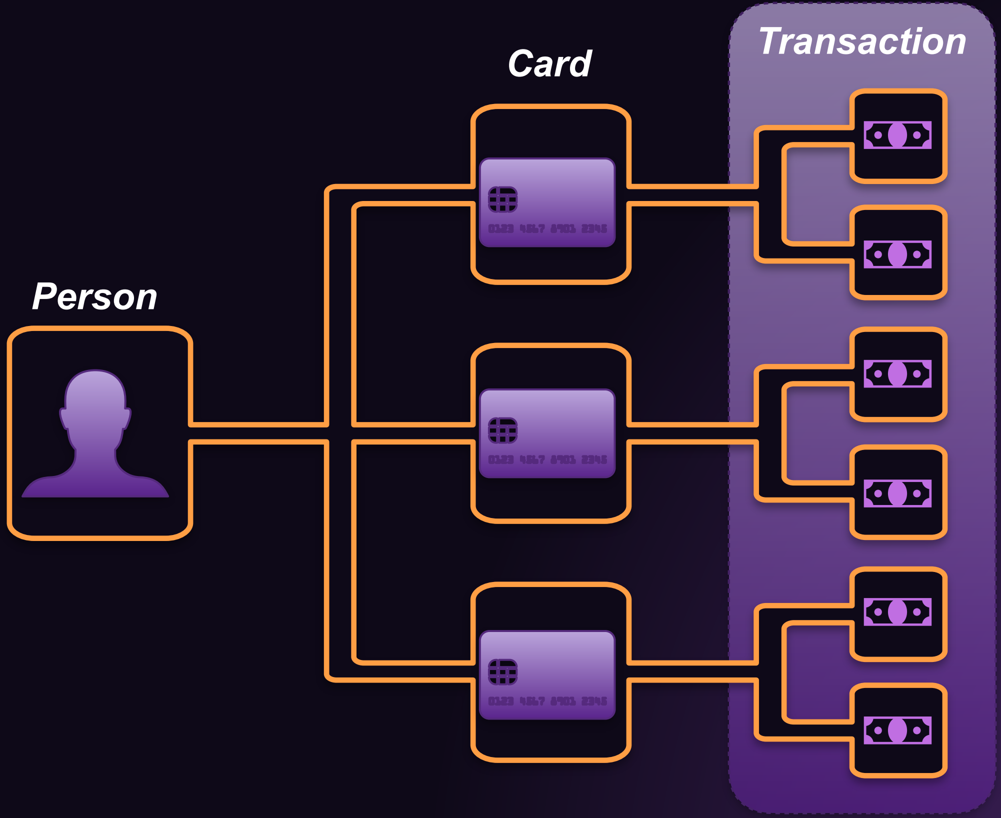
Contents

Fraud Detection & Classification

- Kaggle: *Predict Fraud vs. Non-Fraud Transactions*
  - E.g. Card owner not present
  - Sponsors: IEEE Computational Intelligence Society & Vesta Corporation
- Imbalanced Dataset
  - ~ 600k Transactions, 400+ Attributes
  - Mix of ‘Identity’ & ‘Transaction’ Data
  - Masked Personal Info



Scope



Initial analysis focused on the transaction level, with future analysis moving into grouping by card and person

Approach

Exploratory Data Analysis

- Data Attribute Analysis (fraud/not fraud)
- Feature Generation
- NAN removal

<u>Maximize both precision and recall to greatest extent possible</u>

Modeling

- Logistic Regression
- KNN
- XGBoost
- Limited columns
- Weighting

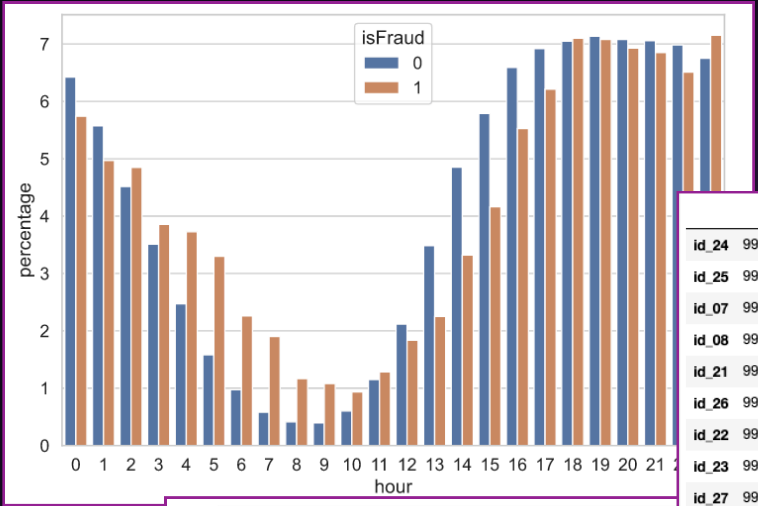
Results & Findings

- XGBoost, Weighted (vs. non-Weighted)

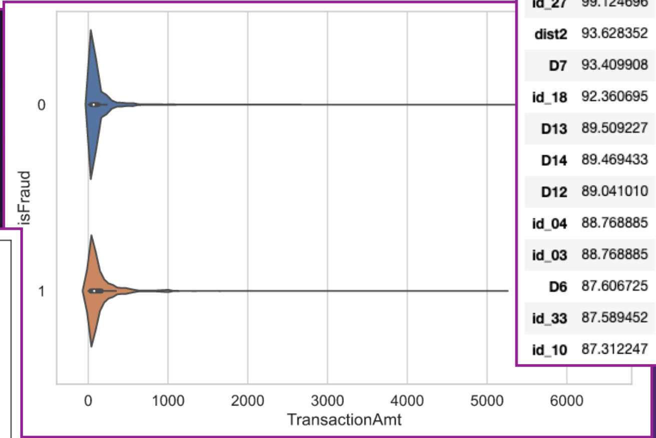
Time, $ amounts and location interaction features emerge as important to the model

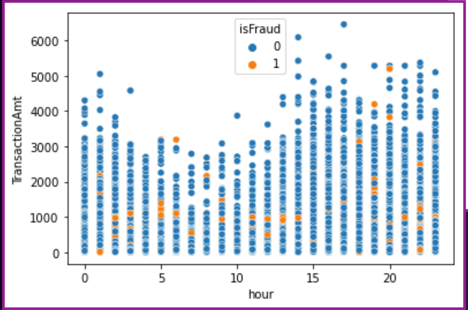
Data Deep-Dive

- Interpret Masked Info
  - e.g. ‘card6’ meaning card type, ‘addr1’ meaning zip or equivalent
- Trim Down Null Columns
- Encode Categorical Features
- Remove Transaction Amt. Outliers
- Add New Time & Address Features



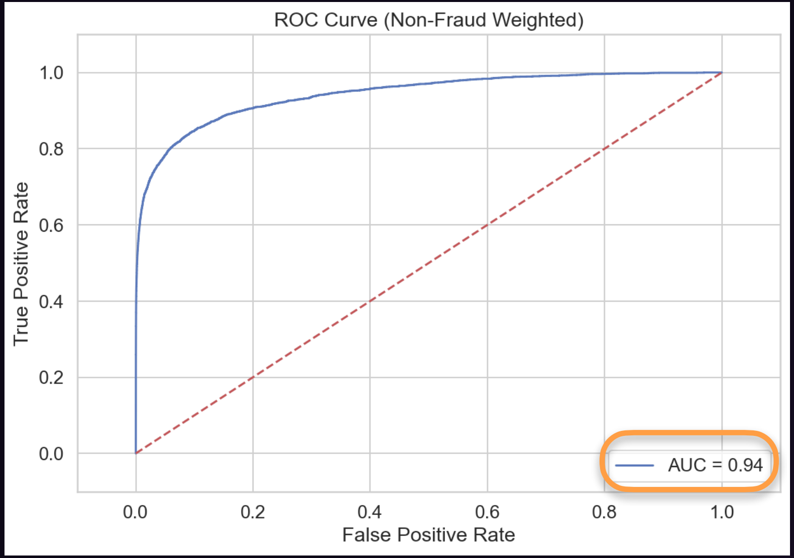
	% Null
id_24	99.196157
id_25	99.130962
id_07	99.127067
id_08	99.127067
id_21	99.126390
id_26	99.125712
id_22	99.124696
id_23	99.124696
id_27	99.124696
dist2	93.628352
D7	93.409908
id_18	92.360695
D13	89.509227
D14	89.469433
D12	89.041010
id_04	88.768885
id_03	88.768885
D6	87.606725
id_33	87.589452
id_10	87.312247

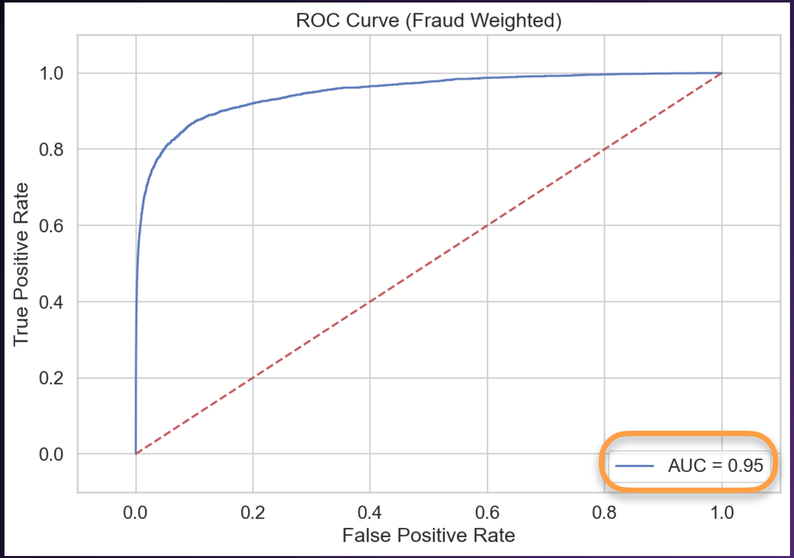




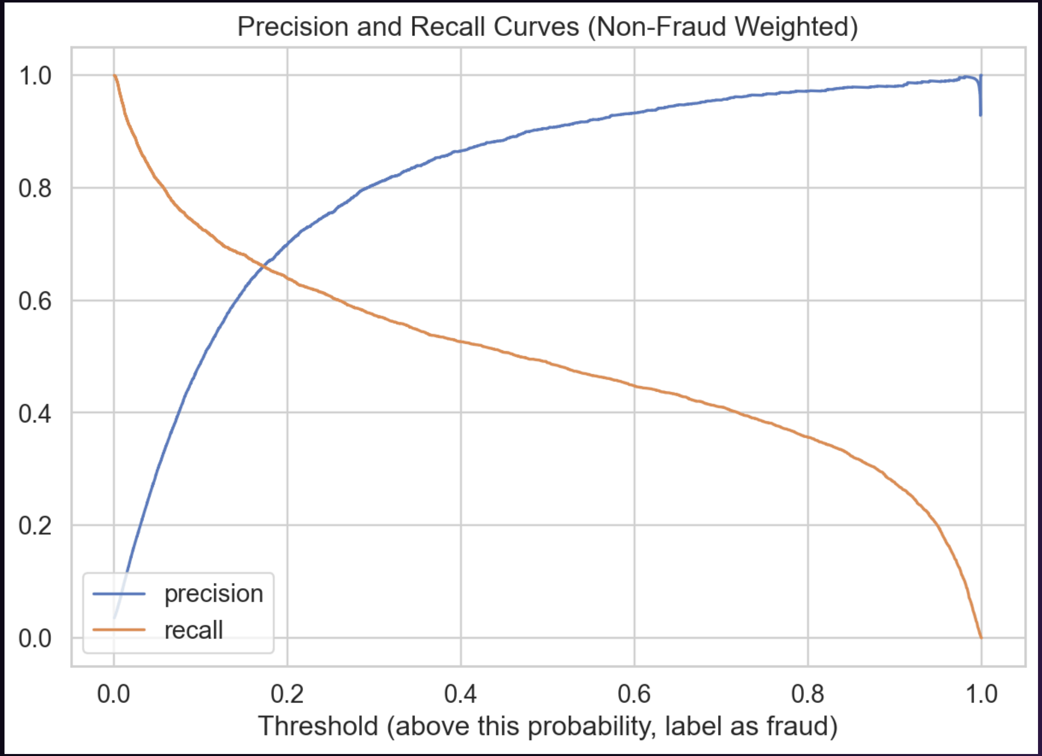
Analysis

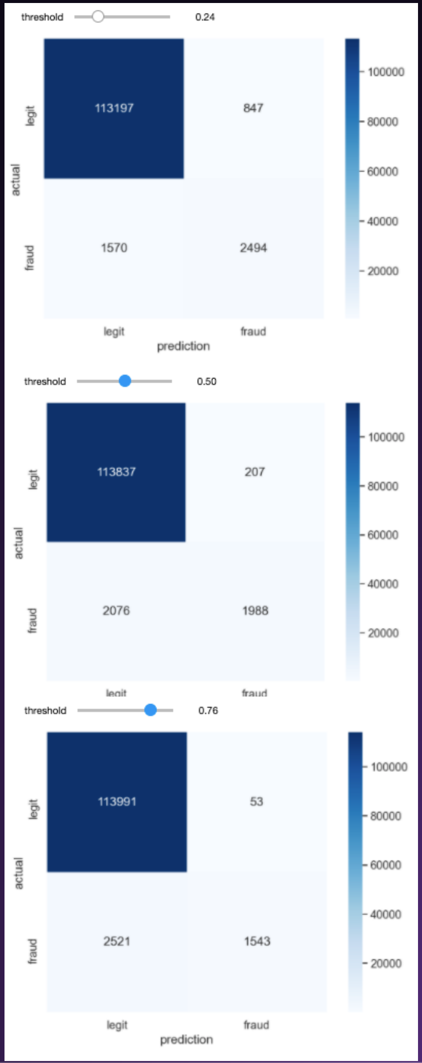
- Larger model with ~200 features
- W/weighted fraud calculations, some AUC+, but minimal



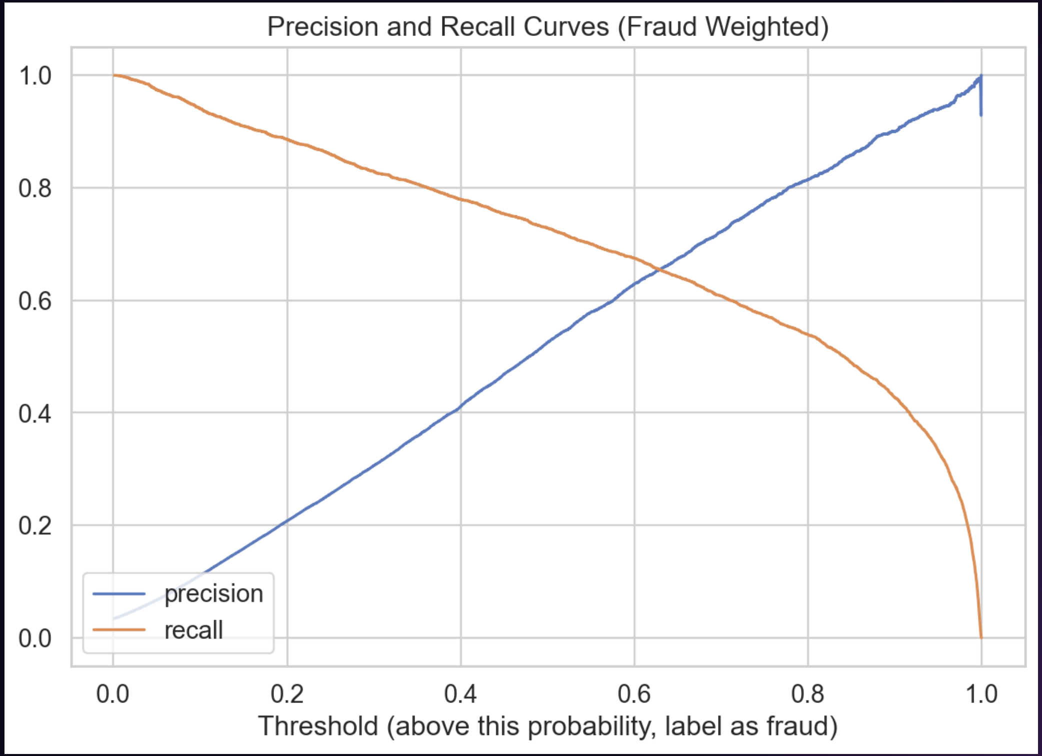


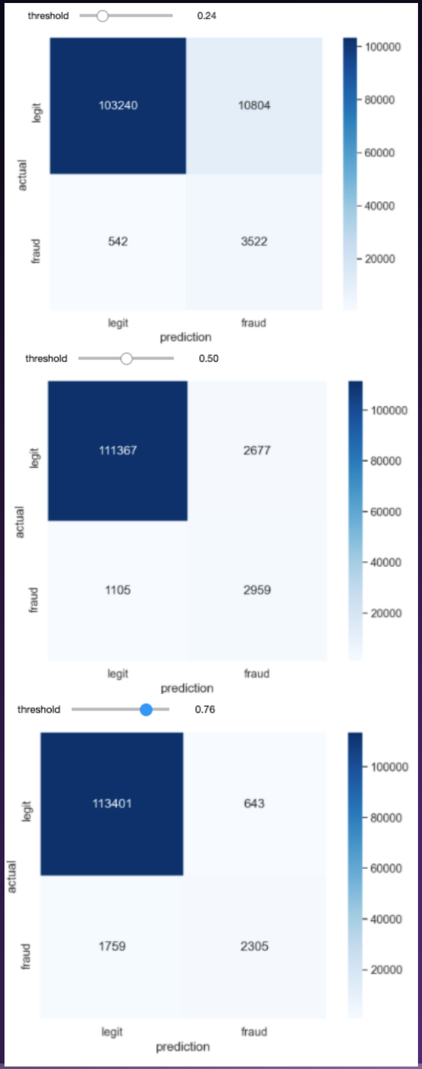
Non-Fraud Weighted





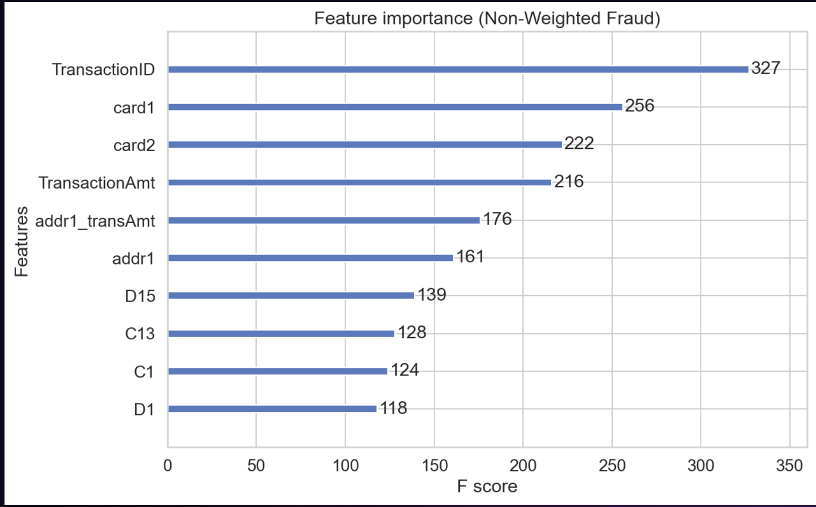
Fraud Weighted

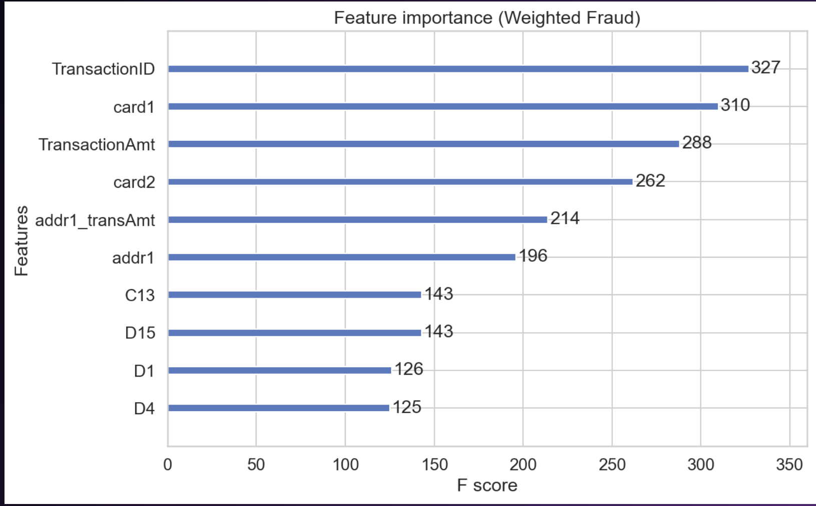




Feature Importance

- C#: Counts (e.g. shared activity)
- Addr#: Location of the purchaser
- D#: Time Related
- V#: Vesta-designed





Key Takeaways & Next Steps

- **Move forward w/weighted XGBoost Model to balance misses and customer impact**

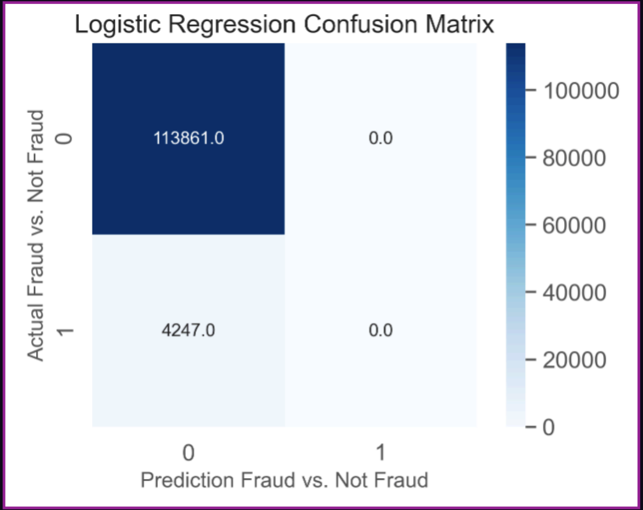
- Continue reviewing features for additional interactions
  (i.e. moving into information on the cardholder)

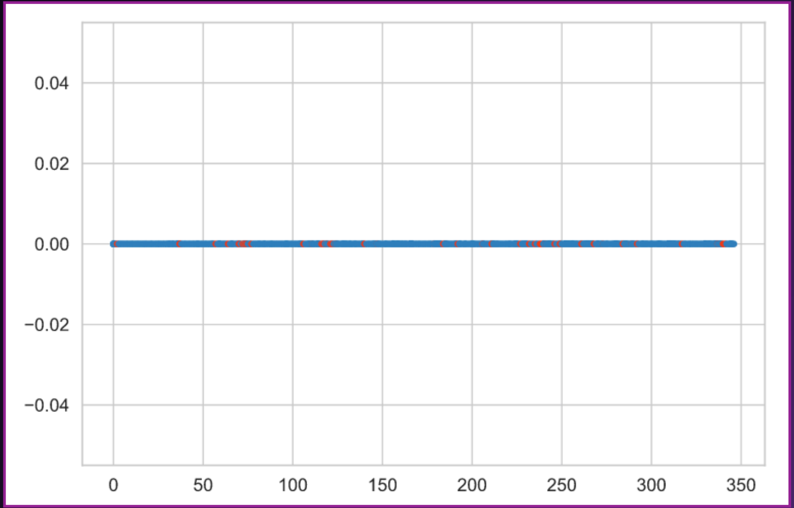
- Identify more duplicative features to further simplify
  (e.g. potentially a number of the created features from Vesta)

Backup

Logistic Regression

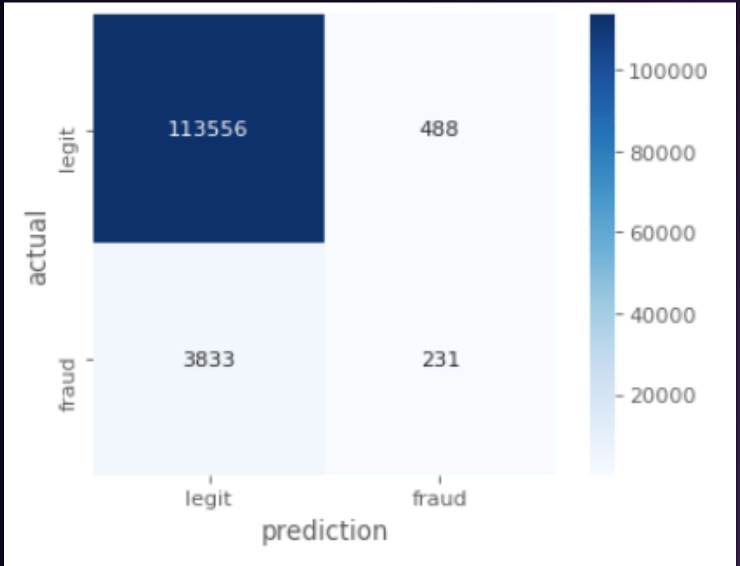
- Poor predictor, based on limited (~20 columns) dataset





K Nearest Neighbor (KNN)

- Improved, but still large False Negative and False Positive Rates



Feature Importance (Gain)