Detecting credit card fraud

IE 5533 Benjamin Lindeen and Michael Ginzburg

Purpose

389,000 cases of credit card fraud per year (2021) - credit.com

The largest data of credit card information affected **160 million cards** in 2009.

Source: U.S. Department of Justice

The FTC had **389,000 reports** of credit card fraud in 2021

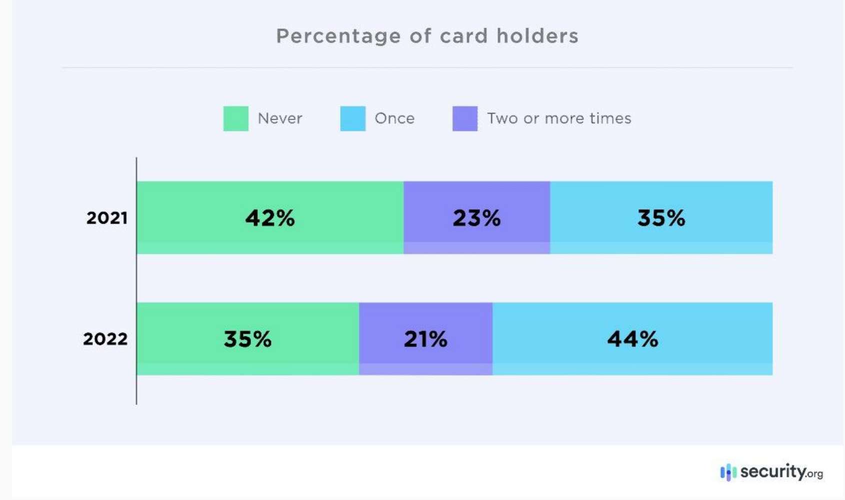
Source: Federal Trade Commission

Purpose

$219 Million of fraud per year in the United States (2022) - Wallethub

Purpose

65 % credit card holders have been fraud victims at some point in their lives - security.org



Purpose

Federal law limits liability to $50 - Investopedia

Companies with $0 liability:

American Express, Bank of America, Barclaycard, Capital One, Chase, Citibank, Discover, PNC Bank, USAA, US Bank, Wells Fargo - NerdWallet

Initial assumptions

Transactions should be seen as suspicious if they are unusually large, or made outside of your normal geographic region.

Examining the Dataset

1 Million total entries

87,404 cases of fraud

Examining the Variables

Distance_from_home (float)

Distance_from_last_transaction (float)

Ratio_to_median_purchase_price (float)

Repeat_retailer (bool)

Used_chip (bool)

Used_pin_number (bool)

Online_order (bool)

Fraud (bool)

Formulating Hypothesis

We presume these variables to be the major contributing factors to credit card fraud: Distance_from_home, Ratio_to_median_purchase_price

First efforts

Linear Regression

Logistic Regression

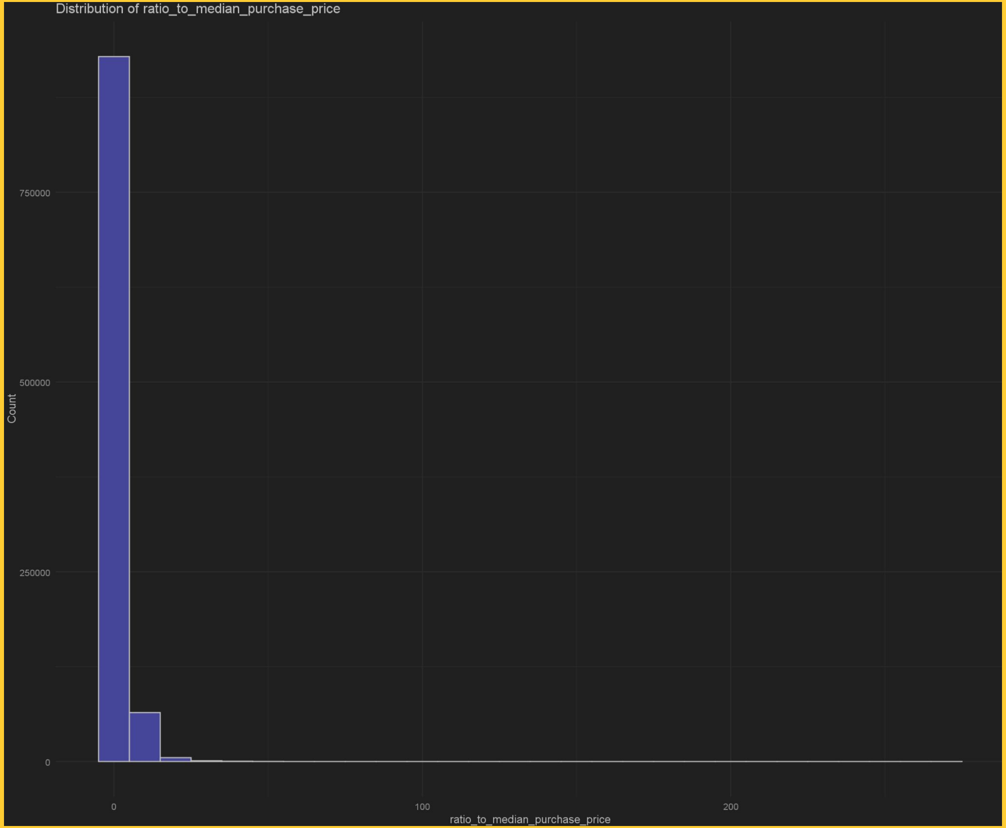
Decision Tree

Finding issues with out approach

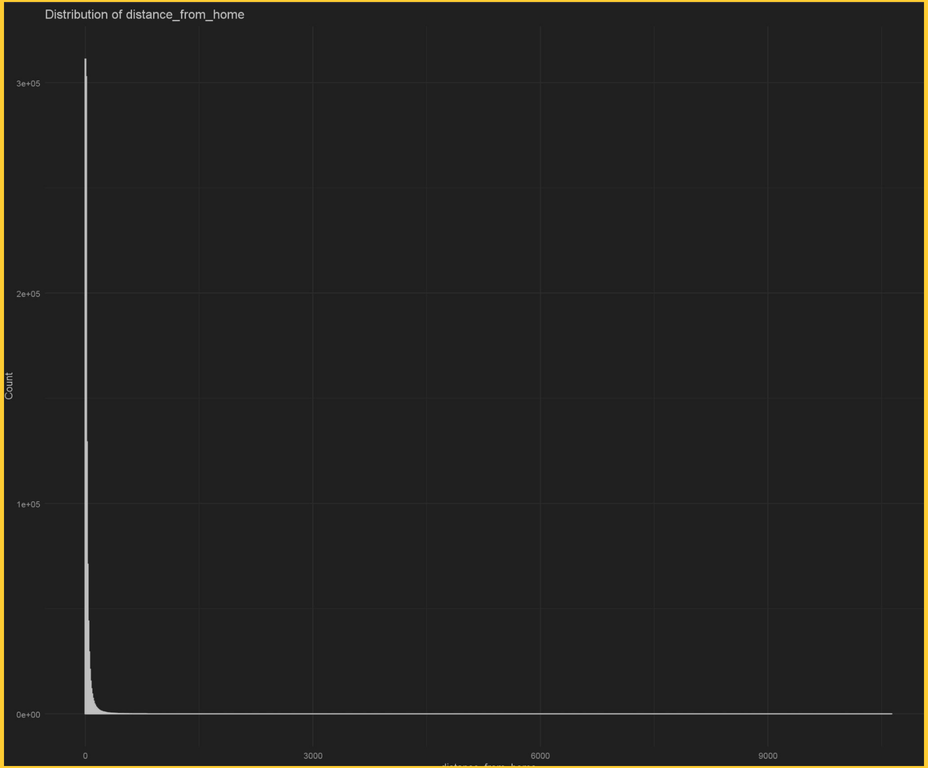
>99% accuracy on decision tree

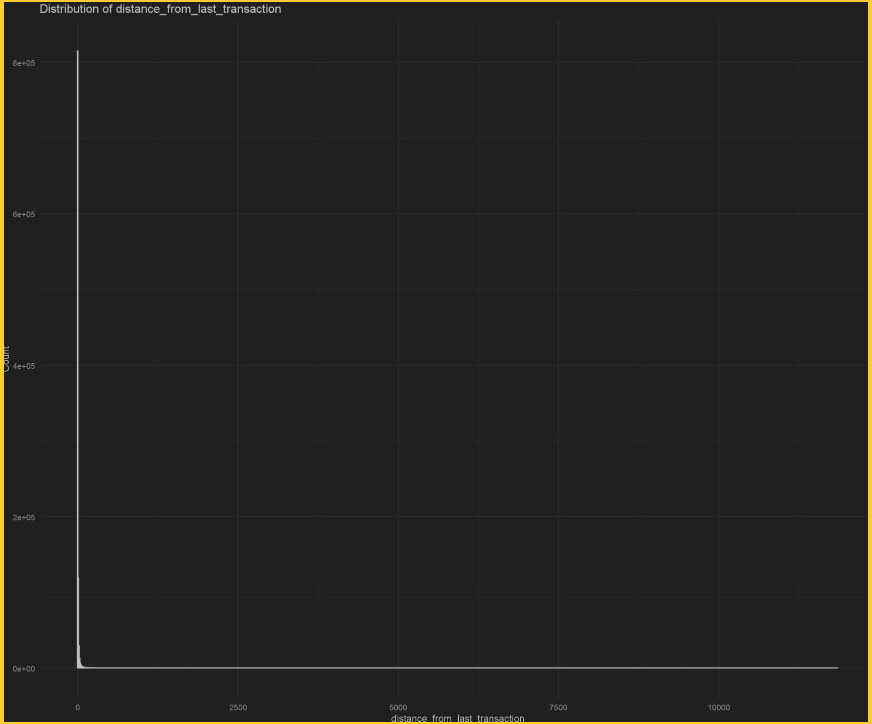
Concluded that our dataset was very imbalance due to the nature of credit card fraud there will be more legitimate transactions than fraudulent ones. A real world dataset will follow this president.

Unbalanced Histograms



Unbalanced Histograms



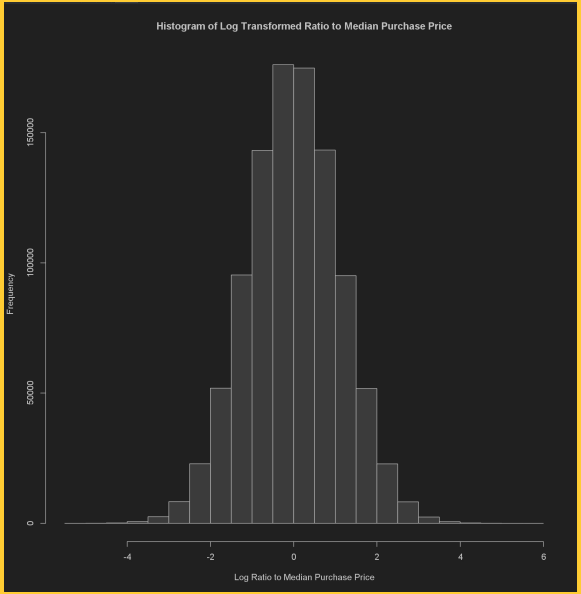


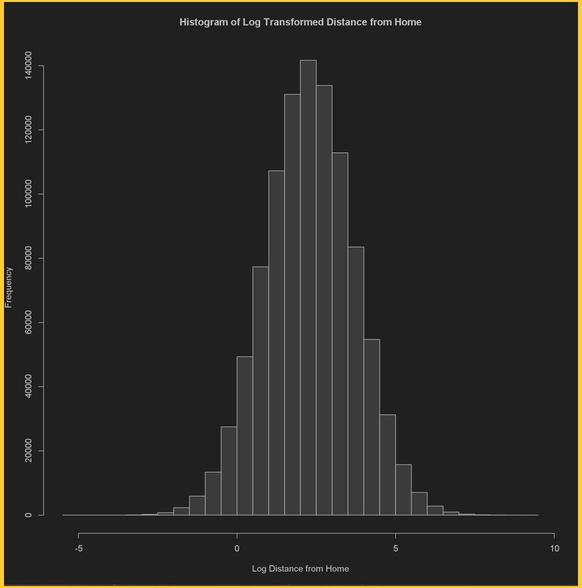
Balancing the data

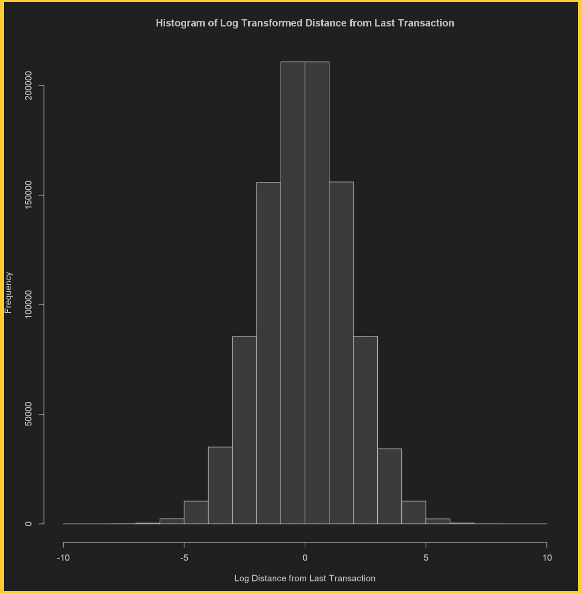
We first started by balancing our data

```
small_constant <- 0.0000001
dataset <- dataset %>%
  mutate(distance_from_home_log = log(ifelse(distance_from_home <= 0, small_constant, distance_from_home)),
         distance_from_last_transaction_log = log(ifelse(distance_from_last_transaction <= 0, small_constant, distance_from_last_transaction)),
         ratio_to_median_purchase_price_log = log(ifelse(ratio_to_median_purchase_price <= 0, small_constant, ratio_to_median_purchase_price)))
```

Balanced data







Revising our strategy

We tried weighing our data.

We tried transforming the data with a log in order to normalize the data.

```
class_weights <- ifelse(training$fraud == 1, (1 / table(training$fraud)[2]), (1 / table(training$fraud)[1]))
```

Success

This worked; Linear model at 82%, Logistic Model at 85% and decision tree at 94% accuracy

Linear Regression

Precision: 0.989563821246989"

Recall: 0.810409532940693"

F1 Score: 0.891070917606662"

	Reference	
Prediction	0	1
0	147920	1560
1	34605	15915

Logistic Regression

Precision: 0.987075918617181"

Recall: 0.84816874400767"

F1 Score: 0.912365483669452"

	Reference	
Prediction	0	1
0	154812	2027
1	27713	15448

Decision Tree

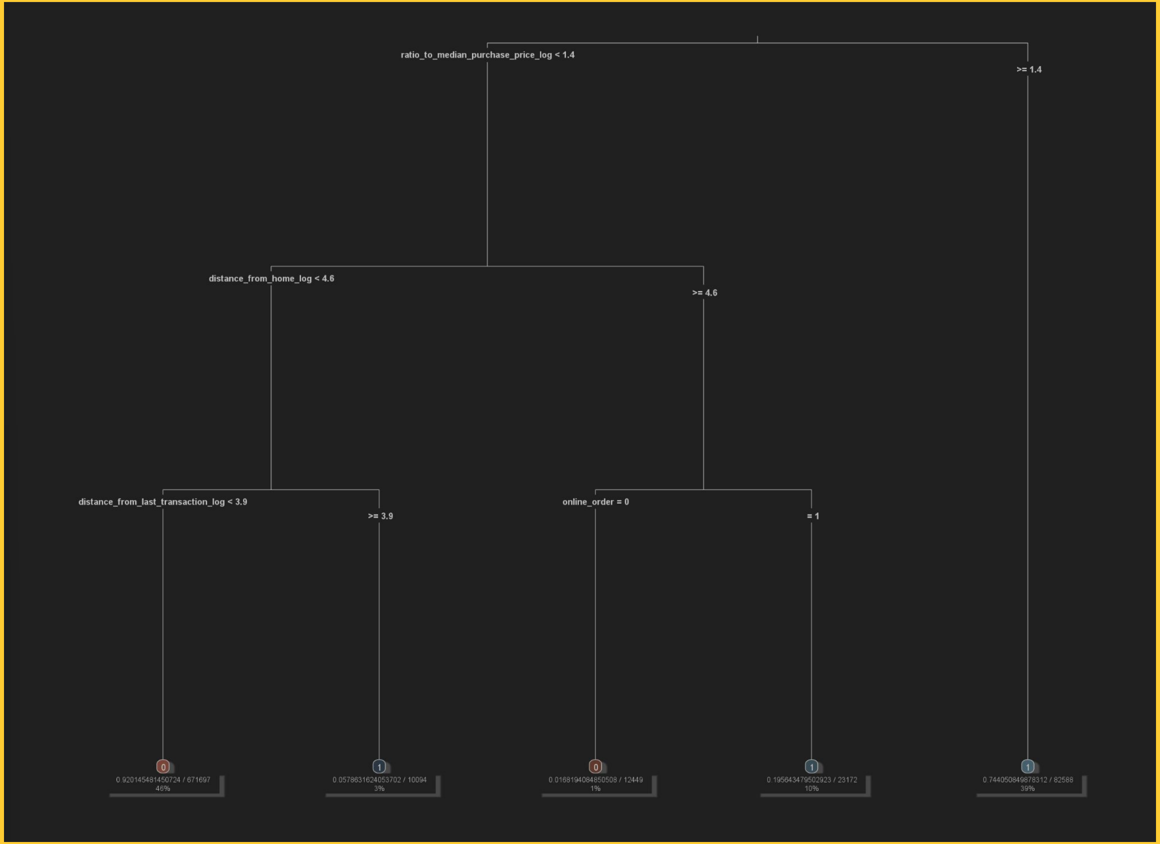
Precision: 0.999766001134895"

Recall: 0.936315573209149"

F1 Score: 0.967001069409788"

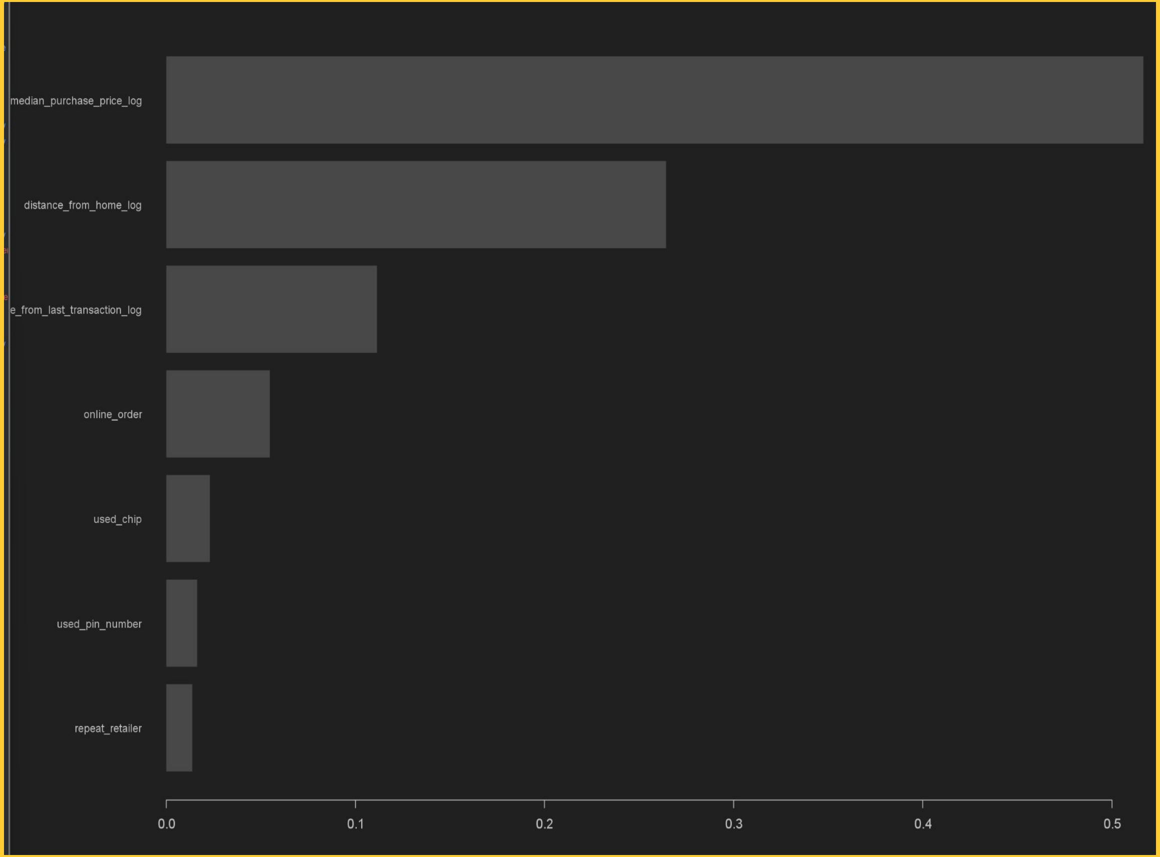
	Reference	
Prediction	0	1
0	170901	40
1	11624	17435

Analyings Results: Decision tree



A high ratio to median purchase indicates fraud, similar purchase price far away from home also indicates fraud.

Analyings Results: Variable importance



Verifying Hypothesis

We were mostly correct, we did not however did not correctly predict that online orders would play as large of a role as they did.

Future potential

- Perform data analysis on only online transactions
  - Presumably more fraud
  - Geographic location much harder to track
  - Random large transactions more common

Thank you!

Questions?