Data Science Immersion Week

Fraud detection

June 2018

Aybike Ulusan
Cristian Gavrus
Yujia Zhou

Mentors:
Frank Ma
Zephy McKanna
Wayne Wang

FRAUD AND WHY WE CARE

- What is fraud?

- Why do we care?

  Wayfair experiences a fraudulent purchase every 4 minutes

  Average fraudulent order costs $719

- How does Wayfair handle fraud?

  Selects suspicious orders and calls customers

  Constraints: limited manpower, affects customer trust

FRAUD $ CAUGHT IS SENSITIVE TO EXPENSIVE ITEMS

$$\underset{\text{(Fraud dollars caught)}}{\text{FDC}} = \frac{\text{Fraud dollar in reviewed orders}}{\text{Total fraud dollar}}$$

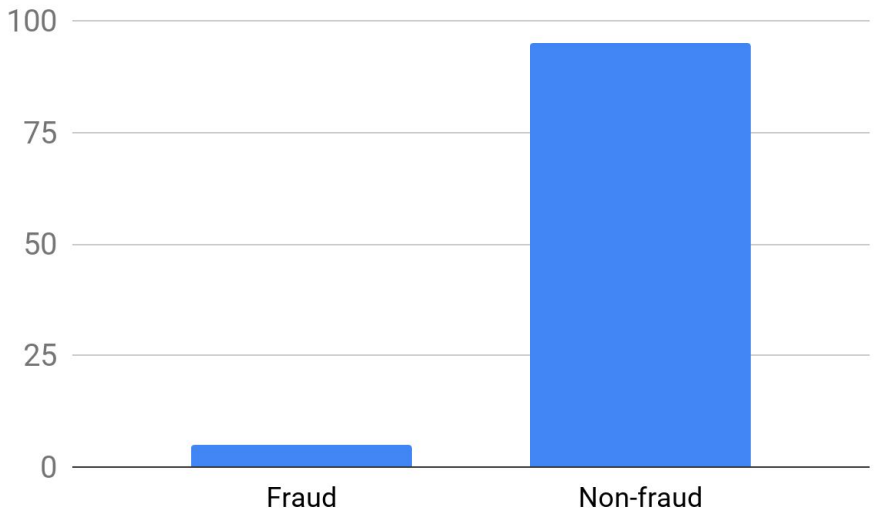
Reviewing and catching a $9 mug

is not as important as a $1300 sofa

FDC IS MORE IMPORTANT THAN PRECISION

$$\text{FDC (Fraud dollars caught)} = \frac{\text{Fraud dollar in reviewed orders}}{\text{Total fraud dollar}}$$

Imbalanced dataset:
5% - 95%



Predicting everything not fraud → 95% precision
But FDC will be 0!

THE GOAL IS MAXIMIZING FDC

Goal:

Maximize FDC within the x orders we submit for review.

DATASET OVERVIEW

Imbalanced

- 100,000 orders (rows)
  - 5% are frauds

Missing data

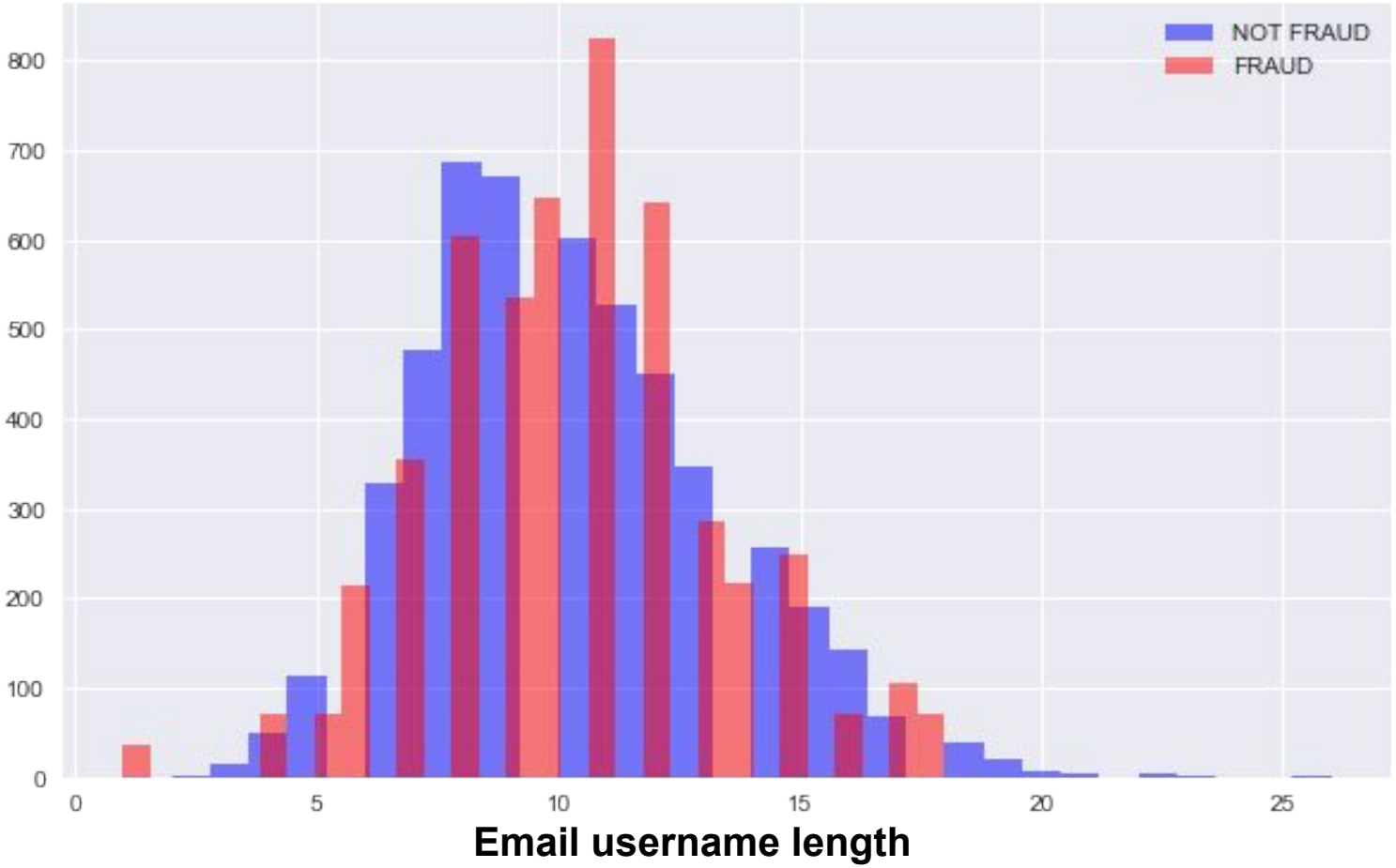
- 106 features
  - Delete columns with >80% missing data (14 features)
  - Median substitution
  - Dummy variables for categorical features

HIGHLY CORRELATED FEATURES

A preliminary analysis identifies two features which should not be used for prediction.

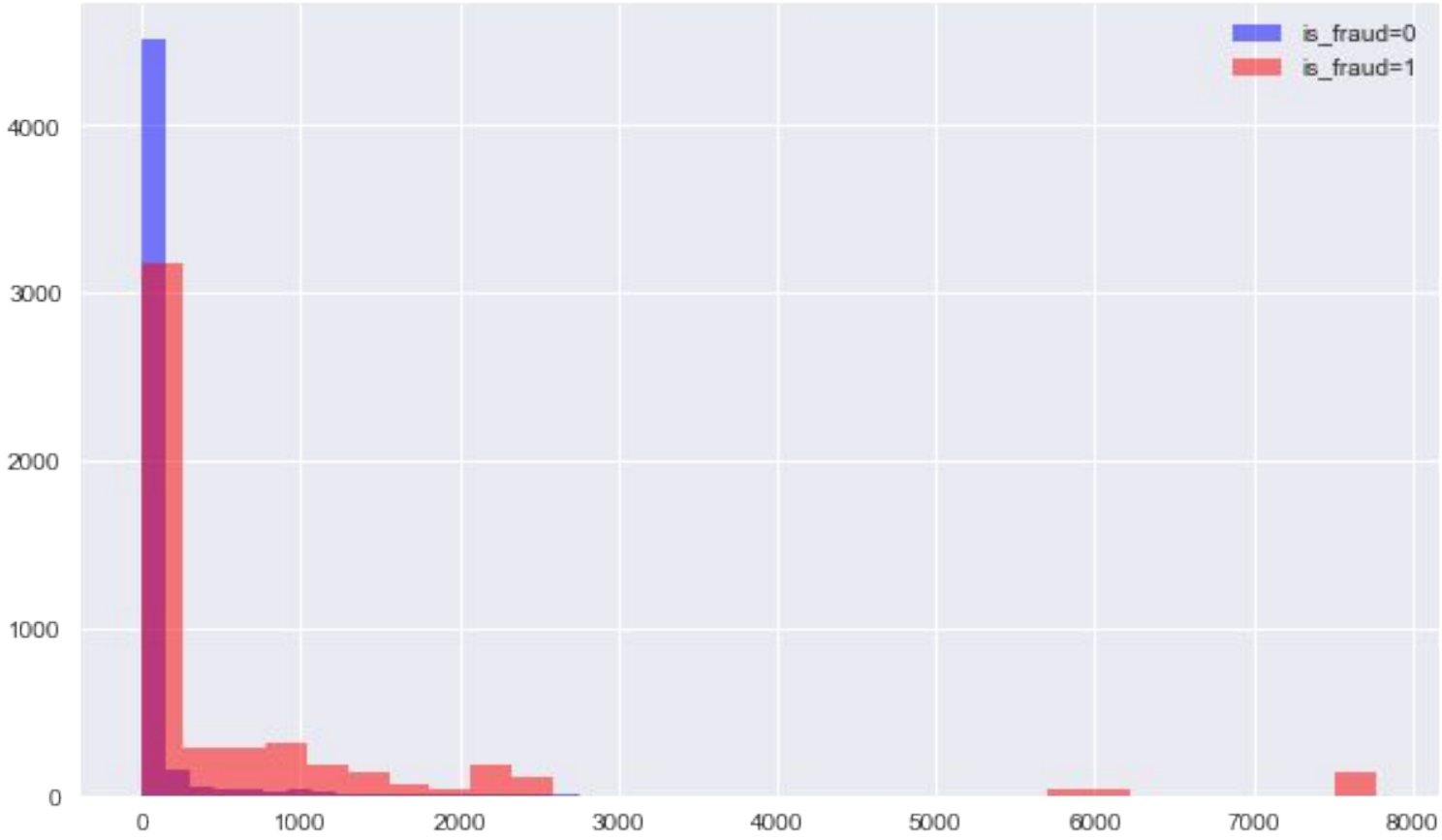
	Correlations: is_fraud
is_fraud	1.000000
is_email_blacklisted	0.986507
is_ip_blacklisted	0.868802
product_blacklist_percent	0.406978
billing_and_ip_distance	0.264345
order_total	0.231995
number_of_orders_in_session	0.224120

EXPLORATORY DATA ANALYSIS - Histograms



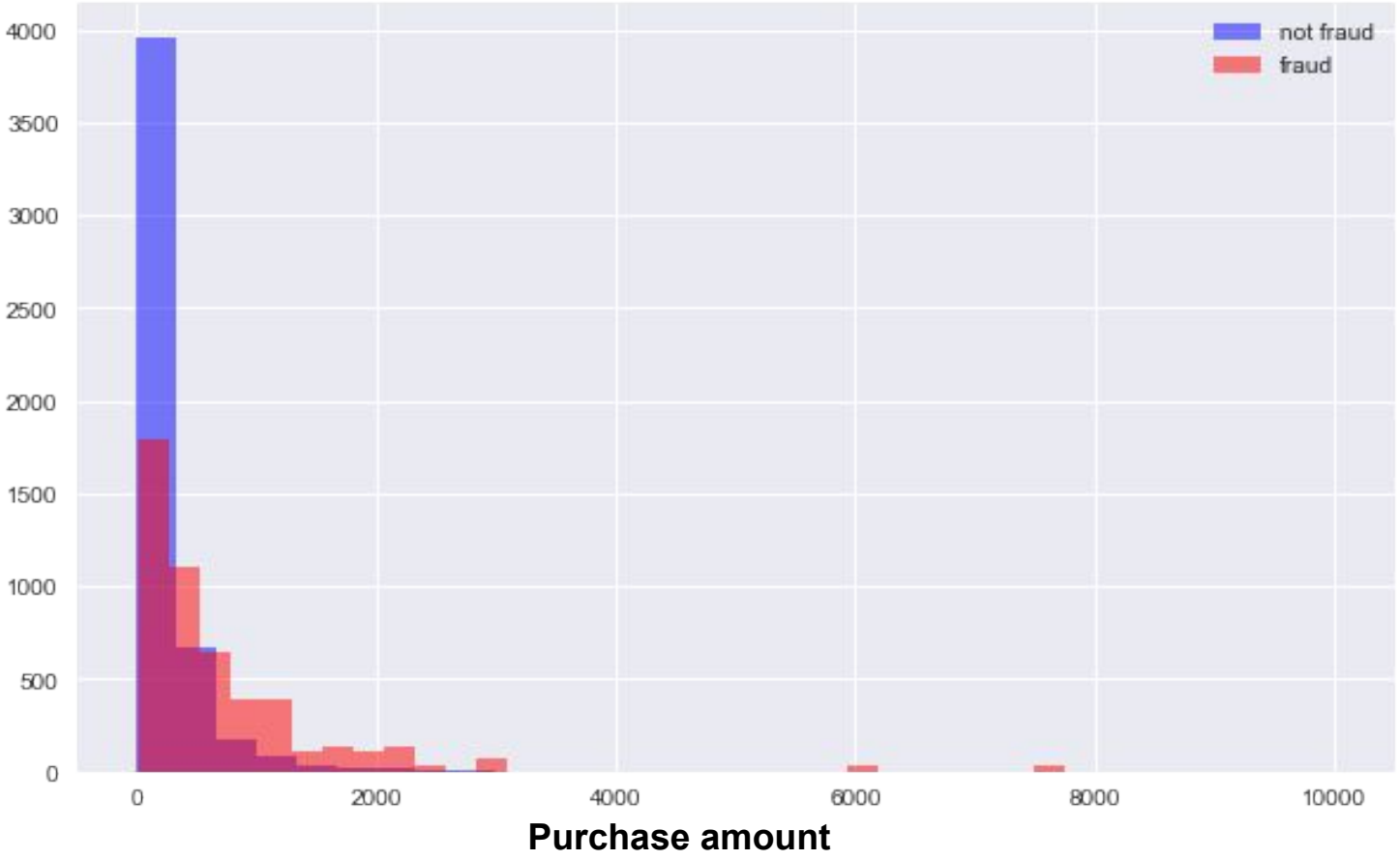
EDA - Billing and IP distance

Fraud cases usually have higher distance between IP and billing locations



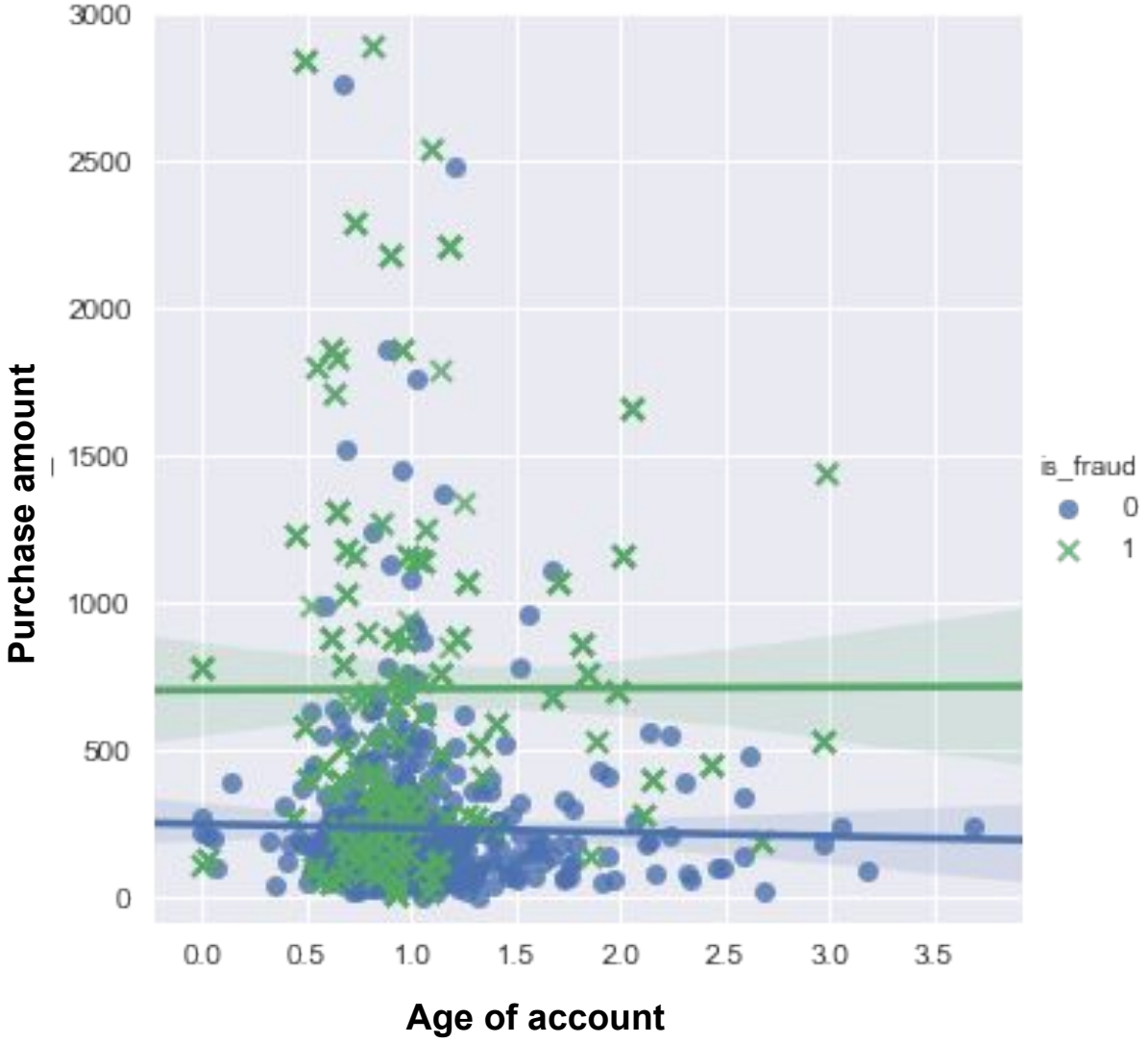
Billing and IP distance

EDA - Purchase amount



Fraud orders spend more

EDA - Purchase amount VS Age of account



New account + expensive order --> more likely to be fraud

Old account + cheap order --> less likely

2-STEP METHODOLOGY: PREDICTION AND RANKING

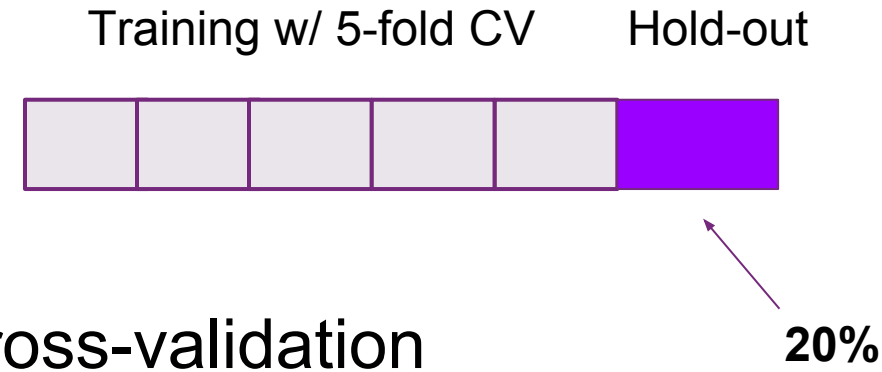
Objective: Maximize FDC (fraud dollar caught)

1. *Binary classification* for predicting the fraudulent activities

2. Rank the cases to maximize $$

OUR ROADMAP

1. Stratified sampling



2. Model selection with cross-validation
   - Logistic Regression
   - SVM
     - linear/nonlinear kernels
     - Option to perform PCA

   Balanced class weight.
   Feature scaling (normalization)

3. Apply model to hold-out set

LOGISTIC REGRESSION

mean of 5-fold CV

Model	Kernel	Features	AUC	FDC%	
				Expected $	Naive
Logistic Regression	Linear	101	0.87	74%	66%

Likelihood	Purchase amount	Expected		
0.9	20	18		x
0.3	500	150	x	
0.6	40	24	x	x

TOP IMPORTANT FEATURES

1. customers_with_same_ip_past_week
2. number_of_orders_in_session
3. number_of_add_to_cart_events_past_month
4. customers_with_same_ip_past_six_months
5. shipping_names_count_past_day
6. shipping_ip_country_match_percentage
7. number_of_keyword_searches_past_week
8. billing_shipping_country_match_percent
9. number_of_pdp_visits_past_week
10. order_total

SVM with LINEAR KERNEL

mean of 5-fold CV

Model	Kernel	Features	AUC	FDC%	
				Expected $	Naive
Logistic Regression	Linear	101	0.87	74%	66%
SVM	Linear	101	0.92	72%	76%

PCA+SVM+NAIVE RANKING YIELDS HIGHEST FDC

mean of 5-fold CV

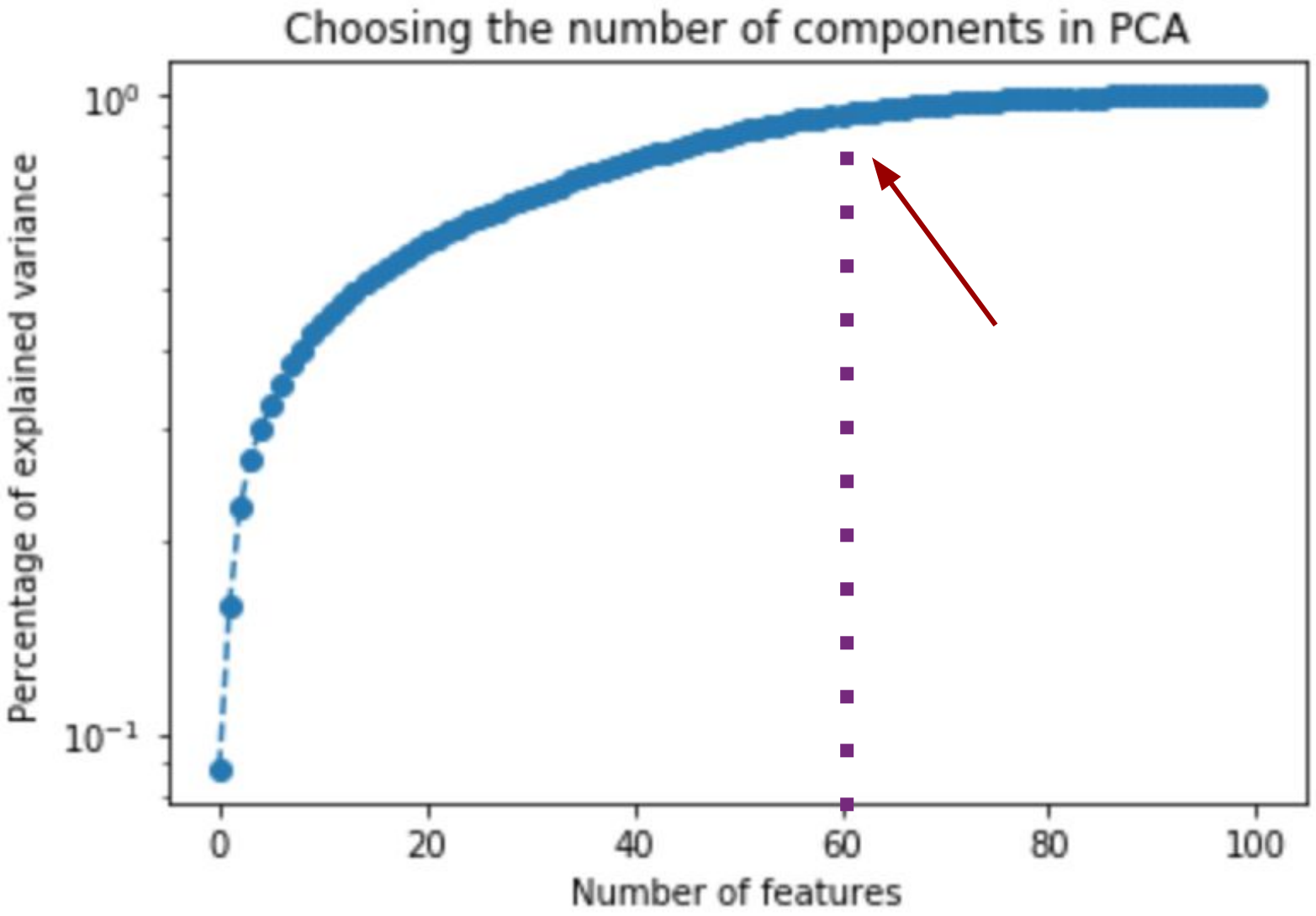
Model	Kernel	Features	AUC	FDC%	
				Expected $	Naive
Logistic Regression	Linear	101	0.87	74%	66%
SVM	Linear	101	0.92	72%	76%
SVM	RBF	101	0.97	86%	93%

PCA+SVM+NAIVE RANKING YIELDS HIGHEST FDC

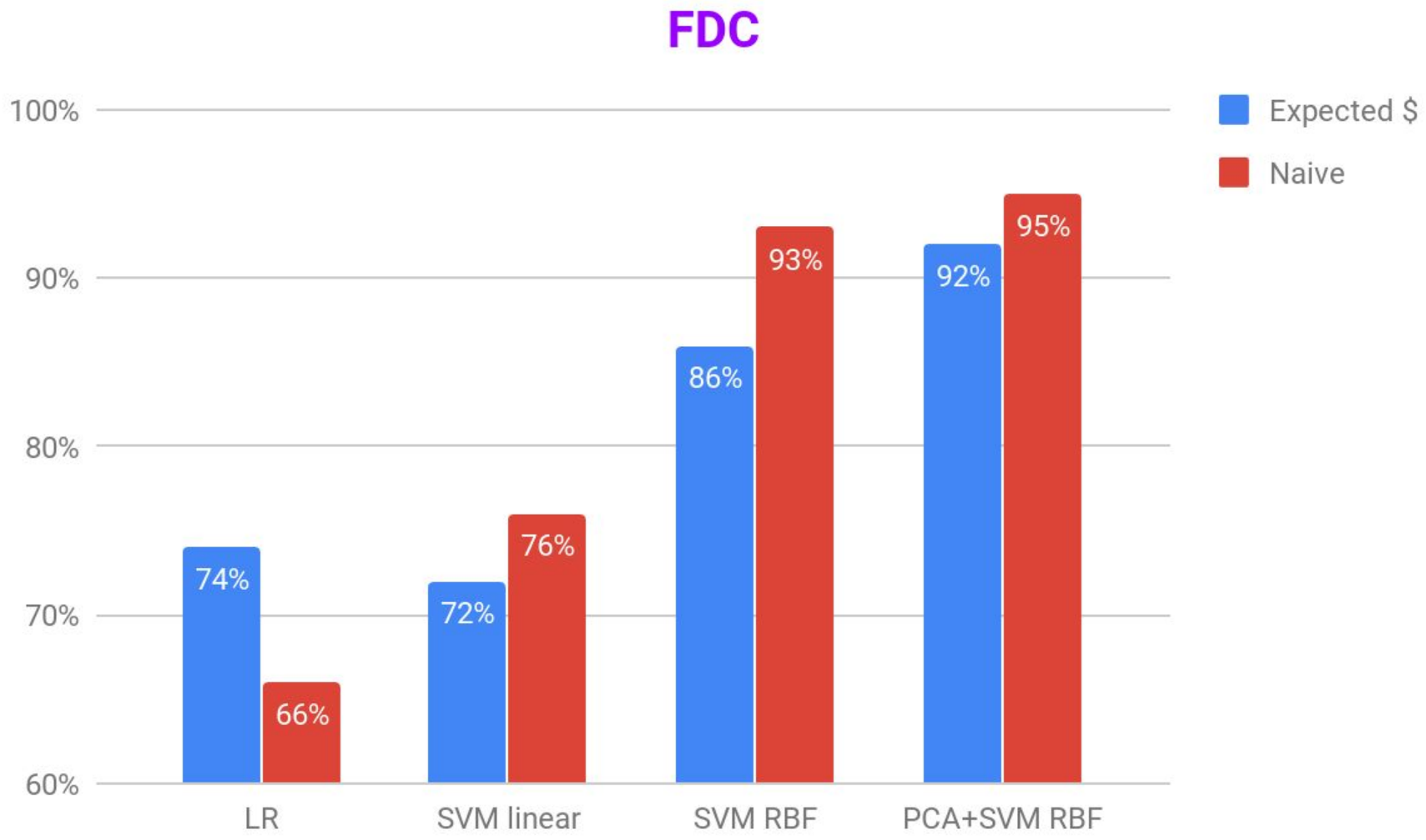
mean of 5-fold CV

Model	Kernel	Features	AUC	FDC%	
				Expected $	Naive
Logistic Regression	Linear	101	0.87	74%	66%
SVM	Linear	101	0.92	72%	76%
SVM	RBF	101	0.97	86%	93%
PCA+SVM	RBF	60	0.99	92%	95%

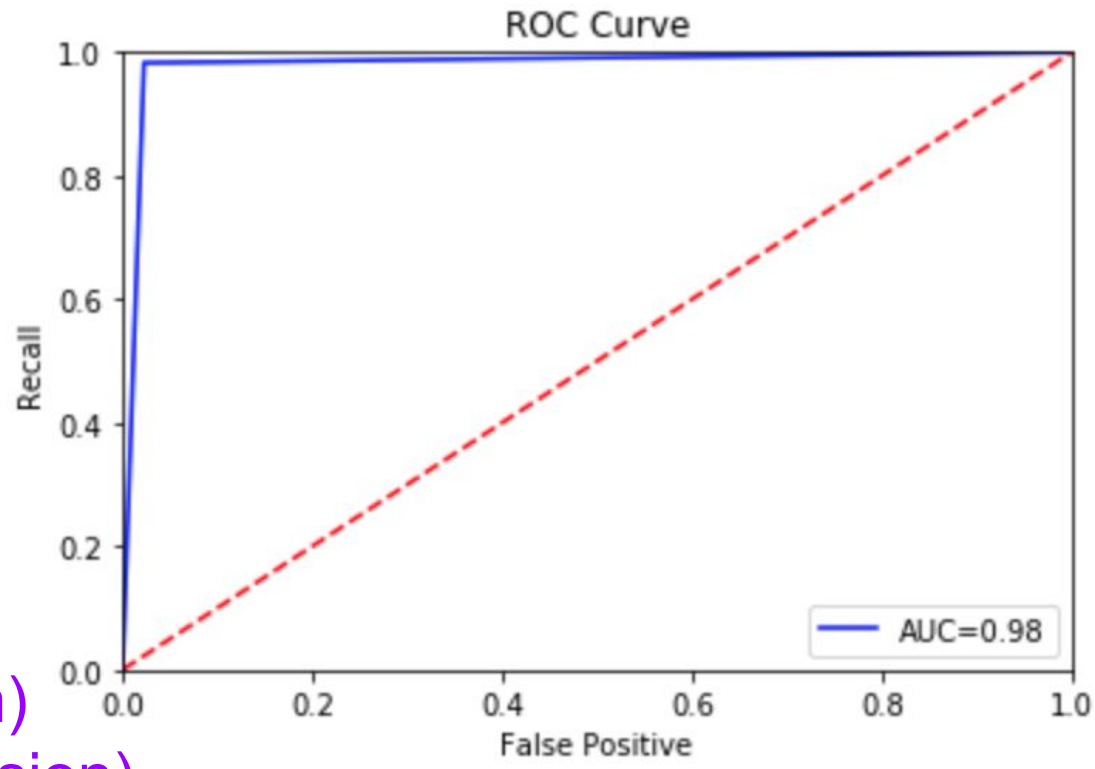
60 COMPONENTS EXPLAINS 93% OF VARIANCE



FDC BY MODEL AND RANKING ALGORITHMS



BEST CLASSIFIER RESULT



Apply to hold-out sample:

FDC = 94% (5% submission)
FDC = 99.8% (7.4% submission)

Preprocessing data with PCA
Pick first 60 components
Fit with SVM (RBF kernel)
Recommend orders to review with naive ranking

FACTORS THAT IMPROVE FDC

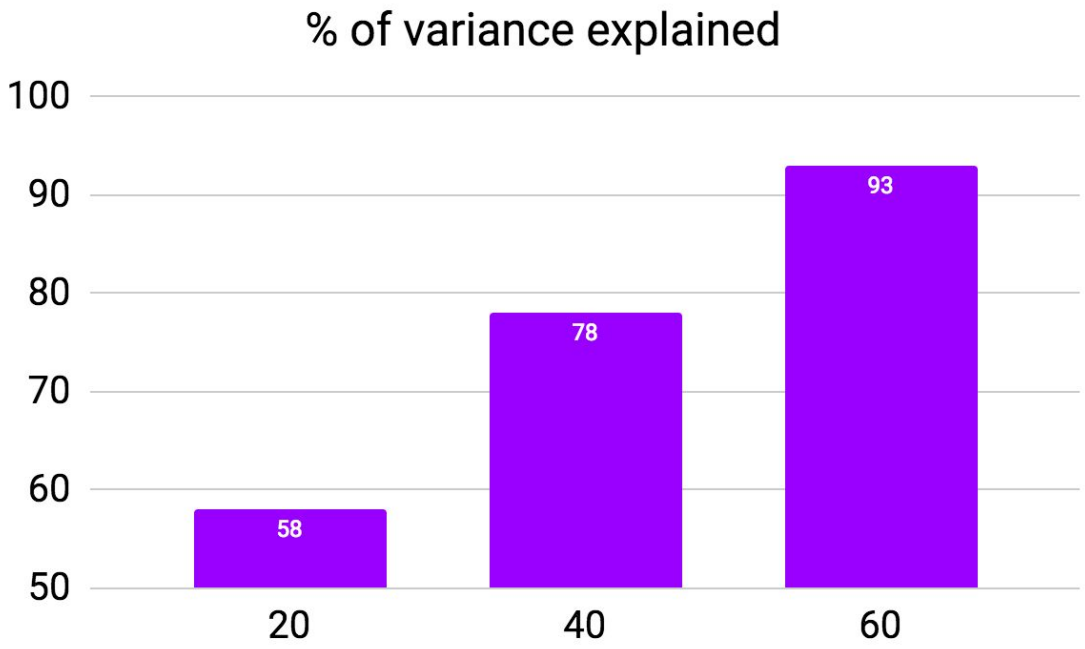
- Nonlinear boundary
  - SVM with RBF kernel performs much better.
- Multicollinearity between features
  - PCA improves the result.
- When predictions are accurate, simply trust the prediction and recommend most expensive orders for review.

FUTURE WORK

- Better ways to handle missing data
- Tune parameters in SVM
- Results too good
  - specific dataset
    - stratified hold-out sample
  - test on “tomorrow’s data”
    - fraud ratio fluctuates day by day

THANK YOU
QUESTIONS?

Other choices for # of principal components



Model	Kernel	Features	FDC%
			Naive
SVM	RBF	20	91.8%
		40	93.1%
		60	94.9%

RANDOM FOREST CLASSIFIER, DO WE TRUST IT?

- In theory, results should be the same on original vs. normalized data.

- In reality,
  - 33% FDC with normalized data
  - 99% FDC with original data

MAP SVM DISTANCE TO PROBABILITY

arctan(x) maps (-∞,∞) to (-π/2,π/2)

f(x) = arctan(x)/π + 0.5