Using Machine Learning to Detect Misstated Financial Statements

Lichen Ni, Leiling Tao

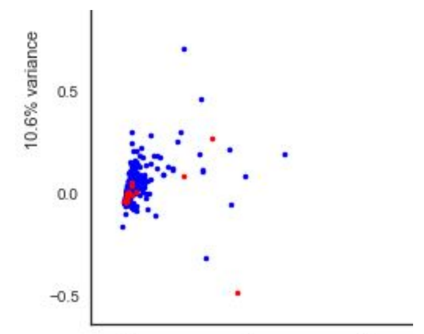
Simon Fraser University

CMPT 733

Introduction

According to the Association of Certified Fraud Examiners, fraudulent financial statements account for 10% of white collar crimes. We aim to automate the process of pre-screening potentially misstated financial statements by using machine learning and interactive visualizations.

Pipeline



- Data source: WRDS - Wharton Research Data Services
- Raw data ranges from Jan 1980 to Mar 2018 with 435497 observations and 981 features

Data Collection

Data Preprocessing / EDA

- Identify misstatement: merge with external data source AAERs - Accounting and Auditing Enforcement Releases
- Extremely unbalanced classes (0.55%)
- Highly correlated accounting variables => need extensive feature engineering
- Focus on a specific country (USA)

Feature Engineering

- Keep the most important accounting terms (5 levels in total)
- Calculating features
  - For ratio analysis
  - According to a paper published by Patricia Dechow for predicting material accounting misstatements

Data Modelling

- Neural networks:
  - Autoencoder
  - LSTM
  - Ensemble neural networks
- Random Forest

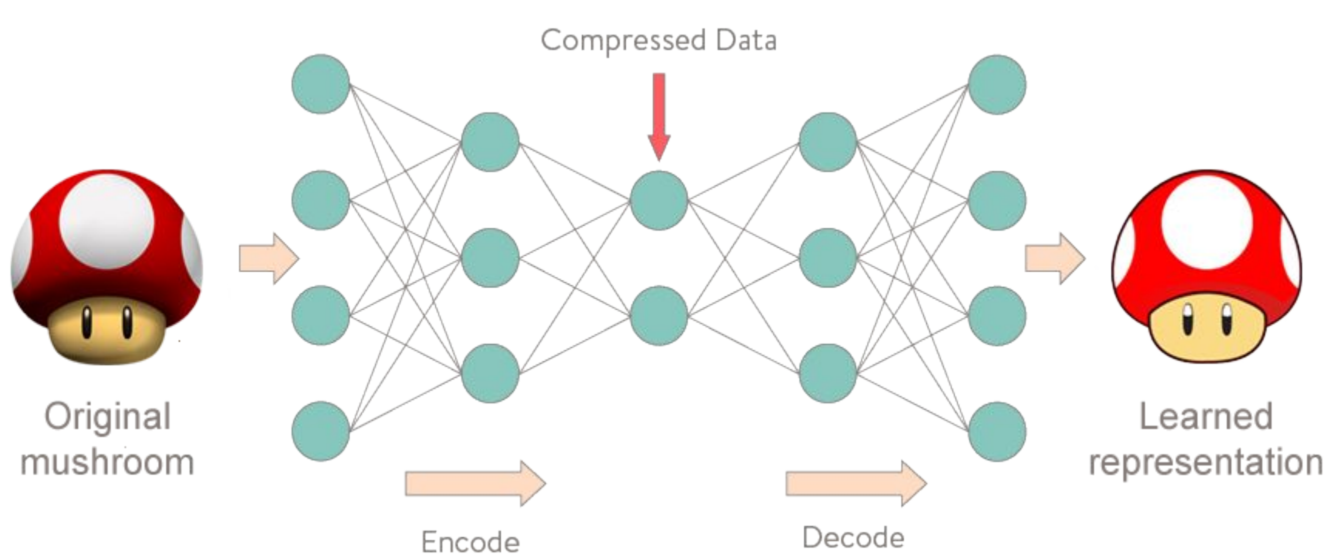
Dashboard

- Incorporating data modelling results
- Use different set of features either form ratio analysis or Dechow
- Targeting a specific industry group in a specific year to identify outliers
- See time series trend for a specific company

Neural Networks

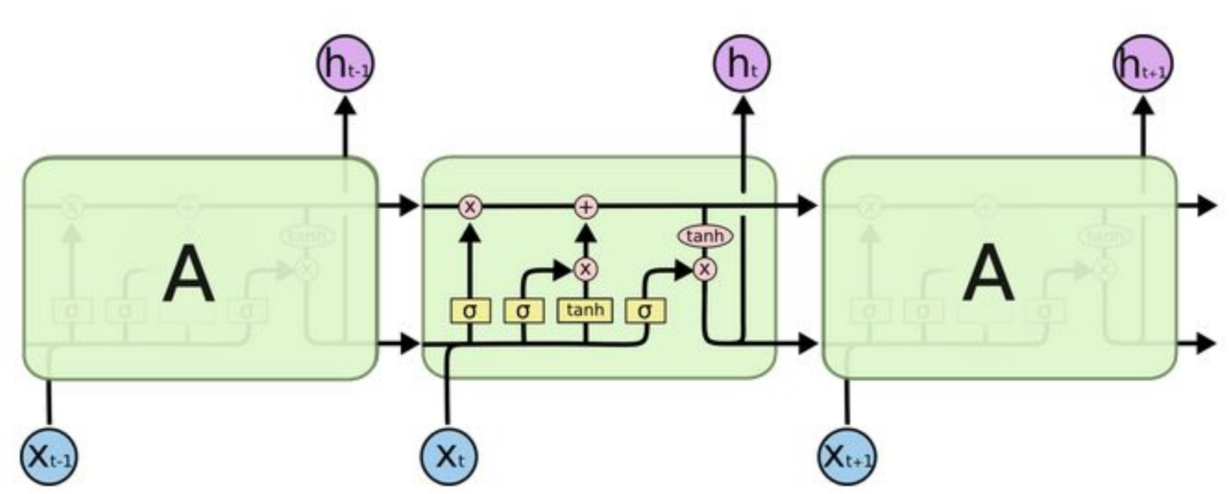
Autoencoder

Finds a compressed identity function of data, which can be interpreted as the underlying structures of the data. Perfect for finding anomalies.



LSTM

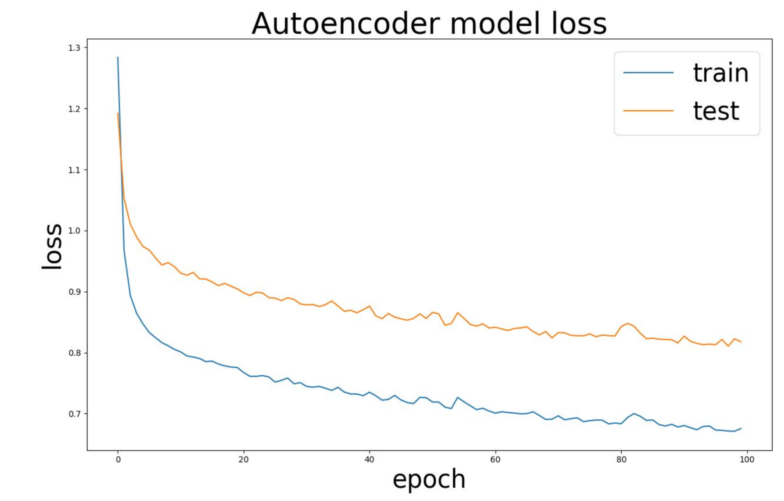
Finds the underlying trend in time. Can predict the data at the next time point; large deviation from prediction means possible anomaly



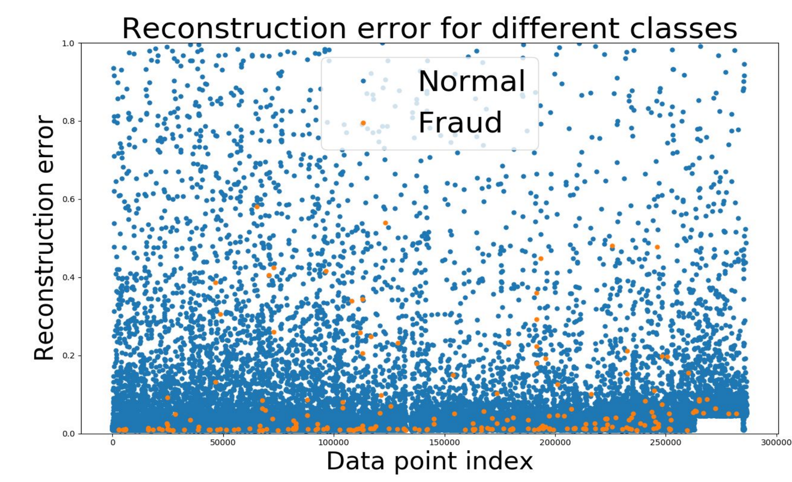
Results

We constructed four neural network models:

- Autoencoder with raw variables
- Autoencoder with calculated ratios
- LSTM with raw variables
- LSTM with calculated ratios



During training, only correctly stated statements were used to learn the underlying structures and time trends. During testing, we made predictions on all testing cases; if the difference between prediction and observation exceeded a threshold, we labeled the case as 'fraud'.



The misstated statements do NOT seem to differ structurally from correct ones (see left).
While the performance of one model was not optimal, we ensembled four models. Compared to a Random Forest Classifier, our meta model had the same precision score but **increased the recall score by 50%**.

Meta-model

True Class	NF	F
NF	45848	6173
F	74	175

Predicted class

Random-forest

True Class	NF	F
NF	50874	6151
F	125	124

Predicted class

Interactive Dashboard