Semi-Supervised Medical Fraud Detection using Autoencoders

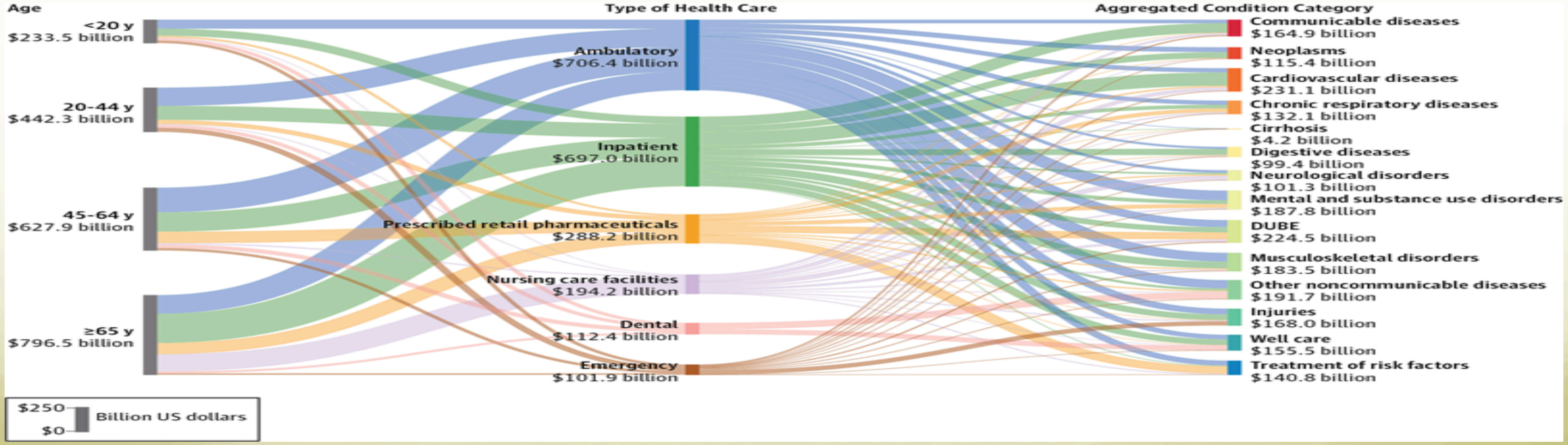
By: Rayhaan Rasheed
The George Washington University
M.S. Data Science Candidate
12/5/2019

Outline

1. U.S. Healthcare Sector
2. Fraud
3. Data
4. Exploratory Data Analysis
5. Anomaly Detection and Autoencoder
6. Data Preprocessing
7. Network Architecture
8. Model Training & Testing
9. Model Evaluation
10. Results
11. Conclusion

U.S. Health Care Sector

- U.S Government spent more than a trillion dollars on it healthcare system in 2018. [1]
- Millions of Americans rely on federally subsidized healthcare to afford procedures, medication, and assistive devices.



(Source: Dielman et. all 2016) [2]

Fraud

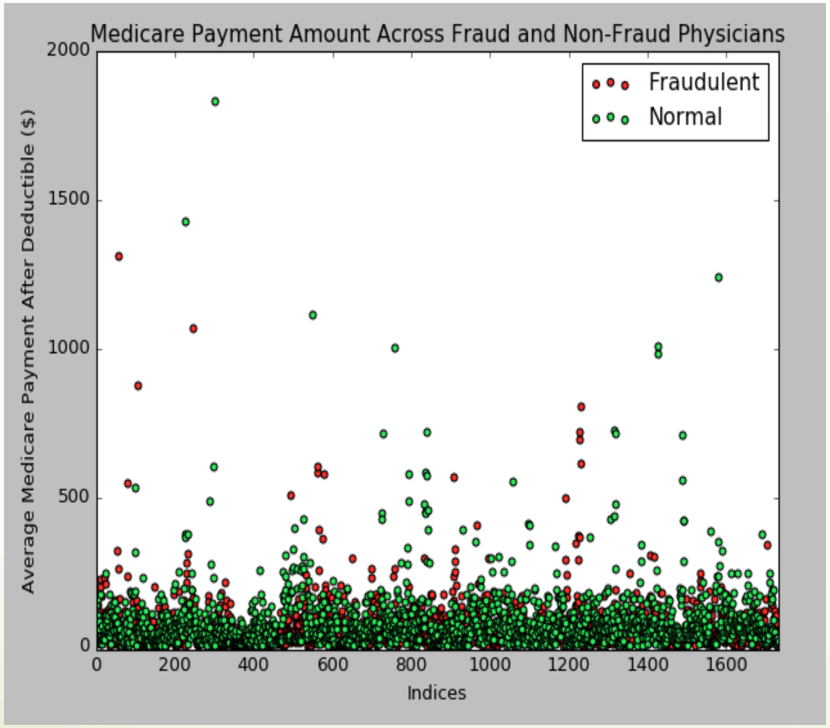
- Instead of billing patient, physicians can get paid from insurance companies and government funds
- FBI estimates more than 10% of total health care spending consists of fraudulent spending [3]
- Types of fraud:
  - Point: Tremendous overbilling for services that were not provided
  - Contextual: Periodic light overbilling
  - Collective: Medically unnecessary procedures

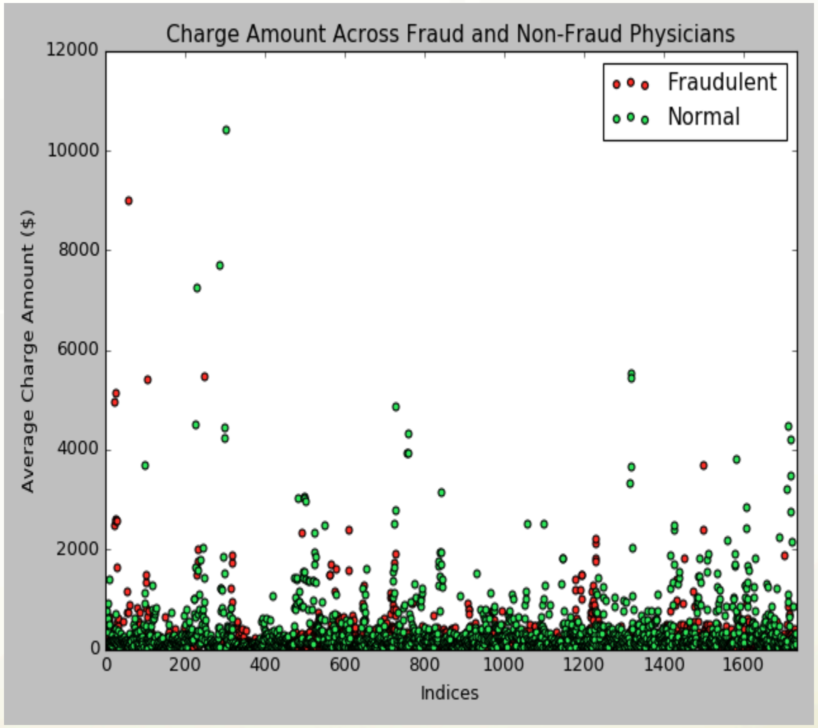
Data

- Provider Utilization and Payment Data (Part B) from the Center for Medicare and Medicaid Services (CMS) [3]
  - Created to aid in detection of fraudulent physicians
  - Procedures performed by physicians with unique NPI code

- List of Excluded Individuals and Entities (LEIE) from the Office of the Inspector General (OIG) at the Dept. of Health and Human Services [4]
  - Exclude anyone from federally funded healthcare programs
  - Each person has a unique NPI code

Exploratory Data Analysis

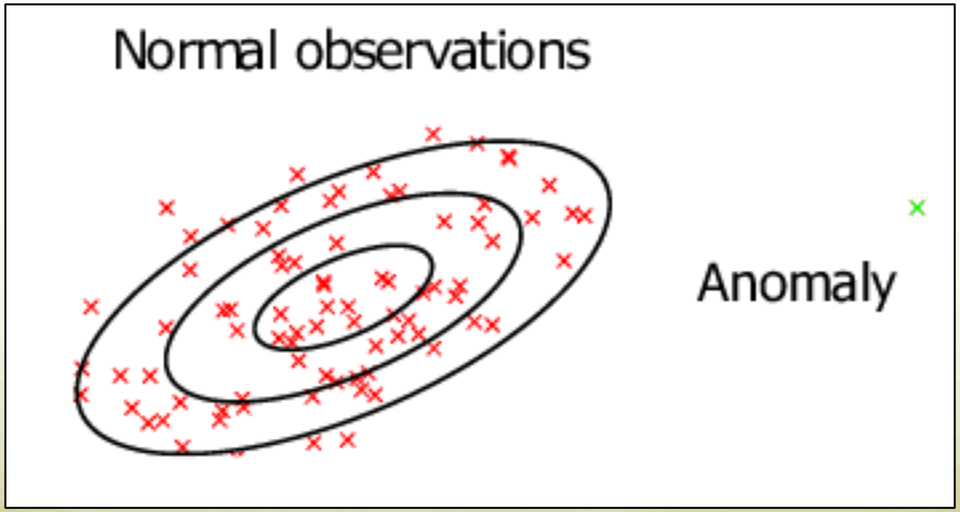




Actual Data: 8,910,479 rows
Normal = 99.9829%
Fraud = 0.0171%

Anomaly Detection

- Given highly imbalanced data, can we detect low-occurring events
- Very different from conventional binary classifiers
- Common AD techniques include Autoencoder, Local Outlier Factor, Isolation Forest, and K-Nearest Neighbor

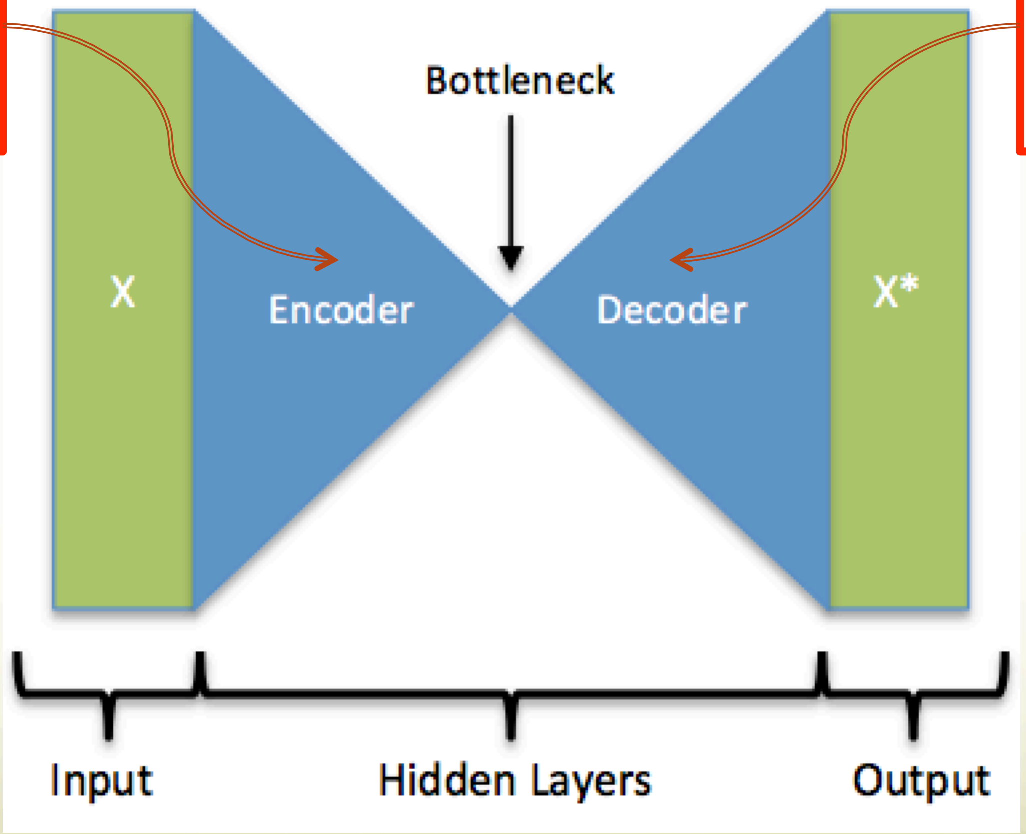


(Source: Assylbekov et. al. 2016)

Autoencoder

Reduce high dimensional input to low-level code

Reconstruct original data using low-level code

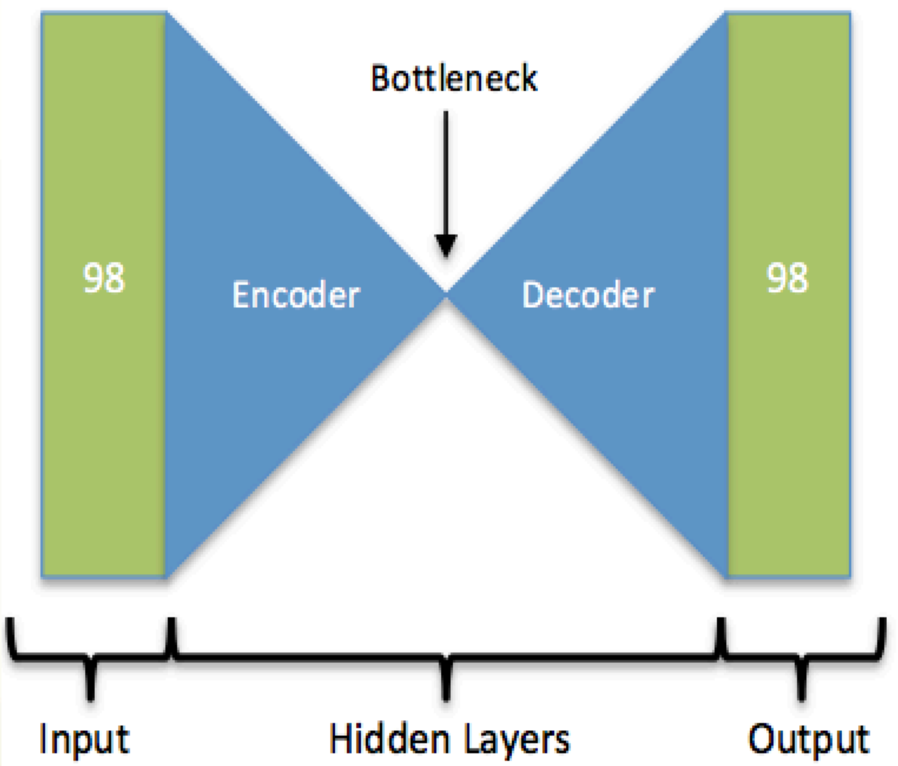


Data Preprocessing

- Preprocessing methodology derived from Herland et. al 2018 [6]
- Add aggregate features for each continuous variable per physician
  - Mean, Standard Deviation, Minimum, and Maximum
- Separate and binarize categorical variables using OneHotEncoder from sciki-learn
  - Drastically increase feature space
- Final, clean dataset contains 8,910,479 observations and 98 attributes

Network Architecture

- Compare three different Autoencoders
  - (98-12-98)
  - (98-35-12-35-98)
  - (98-48-24-12-24-48-98)
- Activation Function: Hyperbolic Tangent
- Train Batch Size = 1000
- Learning Rate = 0.01
- Mean Squared Error (MSE) Loss Function
- Adam Optimizer



How important is network depth given constant hyper parameters?

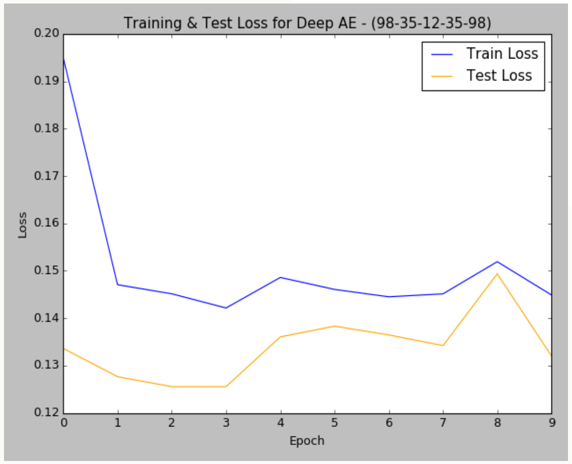
Model Training & Testing

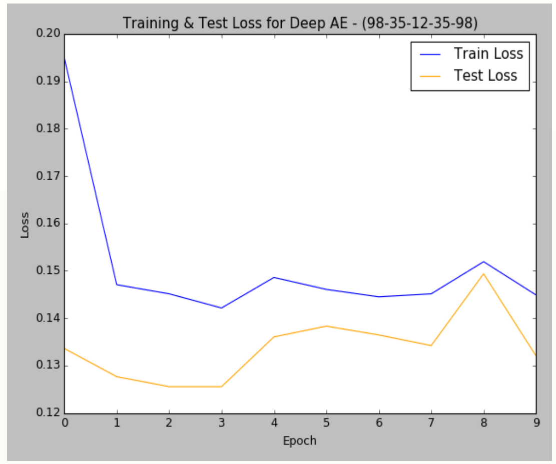
- Semi-Supervised Learning
  - Overall approach is still unsupervised learning
  - Training set is only the non-fraudulent physicians
  - Allows the Autoencoder to learn what "normal" is
- Reconstruction Error (RE)
  - Difference between the reconstructed output and the original input
  - Ideally, fraudulent cases will have a higher RE than normal ones
  - Set discriminatory threshold to classify each physician

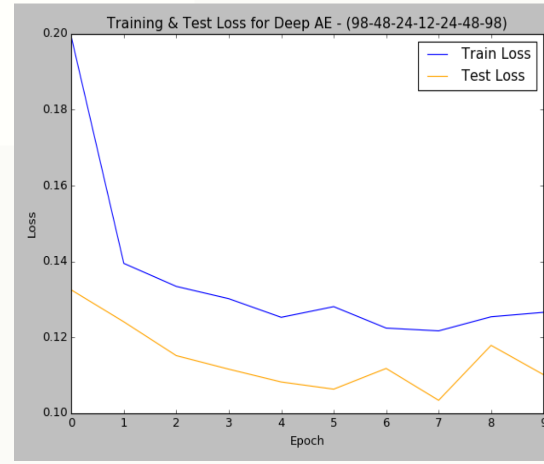
Model Evaluation

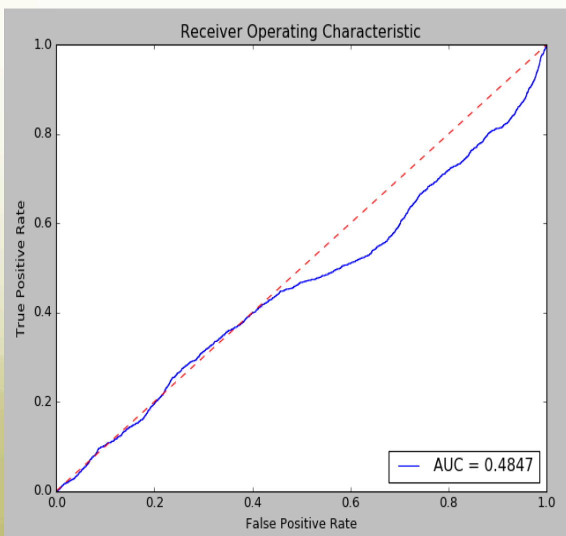
- Recall Score
  - How many actual fraudulent cases does the model capture when we predict fraud
  - A higher recall is good in this case
  - Inflates the false negative value
  - Rather be safe than sorry
- Receiver Operating Characteristic (ROC) Curve
  - False Positive Rate vs. True Positive Rate
  - Area under the cure (AUC) represents overall model performance

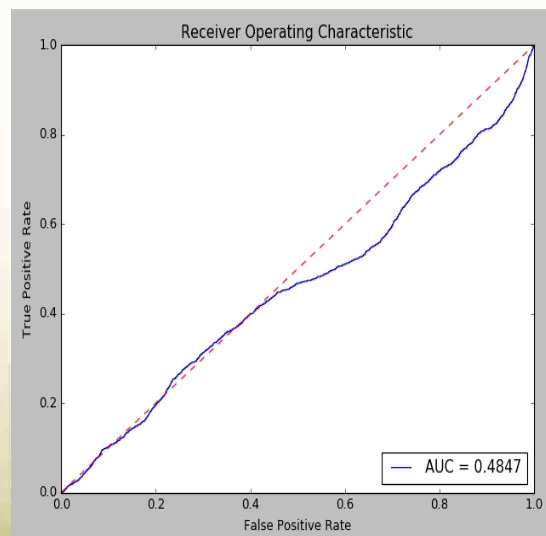
Results

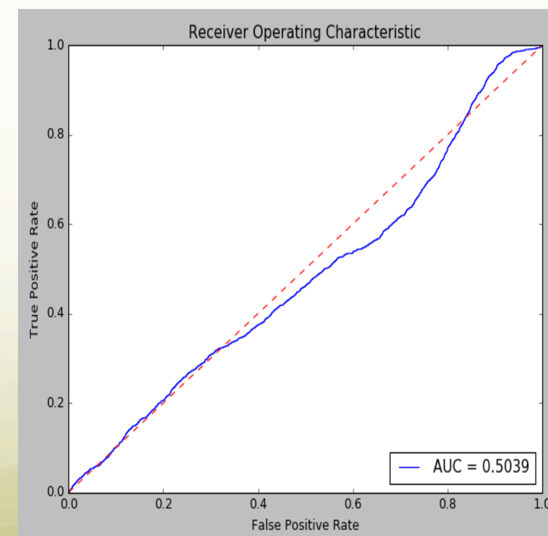

Training & Test Loss for Deep AE - (98-35-12-35-98)
Train Loss
Test Loss
Loss
Epoch


Training & Test Loss for Deep AE - (98-35-12-35-98)
Train Loss
Test Loss
Loss
Epoch


Training & Test Loss for Deep AE - (98-48-24-12-24-48-98)
Train Loss
Test Loss
Loss
Epoch


Receiver Operating Characteristic
AUC = 0.4847
True Positive Rate
False Positive Rate


Receiver Operating Characteristic
AUC = 0.4847
True Positive Rate
False Positive Rate


Receiver Operating Characteristic
AUC = 0.5039
True Positive Rate
False Positive Rate

Results

	Model 1	Model 2	Model 3
AUC	0.5084	0.4847	0.5039
TP	772	706	430
TN	1,351,308	1,350,535	1,938,167
FP	1,311,836	1,322,609	734,977
FN	749	815	1,091
Average Fraud RE	0.7617	0.0444	0.1012
Average Normal RE	0.8433	0.1320	0.04025
Threshold at 50% Recall	0.23	0.03	0.03

Conclusion

- Healthcare fraud is a serious issue that affects the lives of so many individuals
- Fraud Detection methods like Autoencoders are necessary for speeding up he detection process and finding patterns that even humans can not notice
- Future Steps:
  - Various types of Autoencoders
  - Change the hyperparameters of Shallow Autoencoder
  - Use more statistical approaches like the Multivariate Gaussian

Resources

- [1] U.S. Government, U.S. Centers for Medicare & Medicaid Services. The Official U.S. Government Site for Medicare. https://www.medicare.gov/. Accessed 10 Oct 2019.

- [2] Dielman JL, Baral M, et al. US Spending on Personal Health Care and Public Health, 1996-2013. JAMA. 2016;316(24):2627-2646.

- [3] Medicare fraud & abuse: prevention, detection, and reporting. Centers for Medicare & Medicaid Services. 2019. https ://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud_and_Abuse.pdf. Accessed 10 Oct 2019
.

- [4] U.S. Government, U.S. Department of Health and Human Services, Office of the Inspector General. List of Excluded Individuals/Entities. https://oig.hhs.gov/exclusions/. Accessed 10 Oct 2019.

- [5] Assylbekov, Zhenisbek & Melnykov, Igor & Bekishev, Rustam & Baltabayeva, Assel & Bissengaliyeva, Dariya & Mamlin, Eldar. (2016). Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan. 10.1007/978-3-319-39630-9_4.

- [6] Herland *et al. J Big Data (2018) 5:29* https://doi.org/10.1186/s40537-018-0138-3.