Capstone Project Proposals

Idea 1 - Medical Fraud Detection:

Proposal:

Medical fraud resulted in approximately 42 billion dollars in losses in 2020* according to the CMS (Center for Medicare and Medicaid Services). Overbilling, billing for services not rendered and unnecessary services and some of the common types of fraud. Medical insurance company LocalKY Insurance (fictional) would like us to build a system for finding physicians that may be practicing fraudulent charging of services. We could then use the model for examining future medical service charges.

Data:

Using publicly available data from CMS, we have records of billing by provider ID number for the period of time from 2015 to 2019.

Additionally, we have from the US Department of Health Inspector General a list of excluded providers who have been found guilty of fraud, arrested on drug distribution charges or other offenses. (updated as of Dec 2021)

By using the NPI (national provider identification number) we can cross reference between charges and providers who have been excluded by the Inspector General's office.

Analysis techniques:

Because we have a classification problem (fraudulent provider or not), possible methods for analysis could include

Logistic Regression
Decision Trees.
Naive Bayes
Support Vector Machines

Result:

Create a predictive model that flags possibly fraudulent providers for further investigation by the LocalKY Ins. fraud prevention department. Cost savings will be driven by determining fraudulent providers, reduced by the cost of inspecting false positives.

*

https://www.cms.gov/newsroom/fact-sheets/2020-estimated-improper-payment-rates-centers-medicare-medicaid-services-cms-programs, opens new window.

Idea 2 - NLP, Clustering + Web-Scraping Fraud Analysis

Proposal:

Bank of Not-America's customers have been experiencing increasing fraudulent transactions and phishing attempts. While most are not successful, the ones that are have very high costs for the bank. They have asked us to make suggestions for points in their online portal where they might flag or message customers who may either be victims of fraud or at risk of being victimized.

Data:

Using web scraping, we will collect text data from Reddit.com/r/PersonalFinance (posts including words such as 'fraud', 'scam', etc) and categorize it by type of fraud.

Analysis Techniques:

Because we have unlabeled data, we will attempt to determine categories using unsupervised learning.

Natural Language Processing
K-Means Clustering
Hierarchical Clustering
Singular Value Decomposition

Result:

The model we create will be used to propose changes to Bank of Not-America's website and mobile app interface, allowing BofNA to communicate with customers who are engaging in behaviors that look similar to our fraud clusters, reducing costs by preventing fraud that BofNA could otherwise be liable for.

Idea 3 - Amazon Product Review Sentiment Analysis

Proposal:

Companies are getting more and more text data from customers about their products, through reviews on their ecommerce site or social media.

With new products, to add them to the recommendation system, we'd like to determine which products are receiving positive feedback. Additionally, we'd like to flag new items which have extremely negative reviews to possibly remove (counterfeit goods, poor quality, malfunctions).

While some of this data is accompanied by some sort of numeric data (stars, like, etc), often it is not. Our goal is to build a text classifier using Amazon product review data which can be used to analyze customer sentiment which does not have accompanying numeric data.

Data:

The dataset is quite large, in total over 233 million reviews. We are going to select 3 subcategories to analyze:
Electronics
Clothing and Jewelry
Home and Kitchen

https://nijianmo.github.io/amazon/index.html

Analysis Techniques:

Supervised Classification approach:
Naive Bayes - Using a bag of words model, we can use Naive Bayes for a sentiment classifier.
Logistic Regression - If we split the data in half (above 3.5 stars and below), we can build a simple positive/negative sentiment classifier.

Result:

With the developed classifier, our business can now use text reviews which do not have accompanying numeric data to decide which new products to recommend to customers. This will reduce returns and increase customer satisfaction by providing better recommendations.