Dark Web Economics Capstone Proposal

DOMAIN BACKGROUND

Cybersecurity is an ever-growing problem in today's world. Companies are hacked, more often than the public is led to believe, and the hacked data is sold on the dark web. Various markets exist on the dark web, one of the most famous being Silk Road, illegal substances, like drugs and weapons, are exchanged on these markets.

Sans impressive user interface, the dark-net markets (DNM) resemble their sophisticated, legal adaptions, e.g. eBay and Amazon. However, the DNMs must address a complicated issue. How does one create trust amongst criminals and thieves? Some sales use escrow account and have developed arbitration practices.

By understanding DNM economics, society can become better equipped in mitigating cybercrime/crime risk.

This project includes many of my interests – cyber security, markets, and anomaly detection.

PROBLEM

The project will be multi-faceted:

1) Identify top sellers in the Dream Market.
2) Identify malicious products. Which products/sellers are scams?
3) Develop the dark web's Kelly Blue Book, i.e. "Morgan's Black Book"
    a. Sellers can obtain a price expectation to better quantify the risk/return of selling illegal merchandise. Do the expected profits outweigh the risk of being arrested?
    b. The price of illegal goods is opaque. The model will create price transparency.

DATASET

AZSecure – a consortium of various U.S. universities – obtained data of ~120k product sales and ~2.5k sellers from the Dream Market for 2016 and 2017.

Dream Market was used to sell stolen credentials that were procured during the Equifax data breach that occurred in 2017.

SOLUTION

1) To understand what leads to the top sellers, I will analyze the sellers' product offerings/descriptions, positive/negative ratings, location, etc. I will define a top seller by either (1) highest total gross sales or (2) customer rating or (3) order of rating

    By accurately predicting any of those metrics, I expect to be able to infer what attributes are shared among top sellers.

    One bottleneck is that I am unsure whether to merge the 2016 and 2017 seller data sets. There are > 300 sellers that are on both datasets and their ratings reflect those moments in time. Also, I expect these sellers to have higher gross sales. Consequently, I am favoring using customer rating or the order of the ratings (descending) as the metric. If I use customer rating, I must determine which rating to associate with the user.

Two potential resolutions:

- Keep the two counts of the sellers. Will this skew the data? Is there such thing as "observation collinearity?"
- Drop the redundant sellers. I don't' want to lose these sellers because I suspect they have many sales and possess loads of quality to be used to train the models.

When I say order, I refer to comparing the model's performance not based on the seller's review score, ranging 0 to 5, but the order in which the model predicts the seller's ratings. I would evaluate this metric by using Spearman's Coefficient.

2) To identify malicious transactions, i.e. scams, I will apply anomaly detection. This will be unsupervised because the dataset does not identify which products were malevolent.

3) To construct the price estimator model, I will develop a regression model, most likely an ensemble.

BENCHMARK MODEL

1) Identify top sellers – Linear Regression and/or Decision Tree

2) Need help in defining what to use as a benchmark model for unsupervised machine learning.

3) Price regression model – Linear Regression and/or Decision Tree

EVALUATION METRICS

1) Identify top sellers – RMSE or Spearman's coefficient

2) Anomaly detection - Silhouette

3) Price regression model – RMSE

PROJECT DESIGN

** There will be iteration between steps 2 and 3.

** Steps 4, 5, and 6 are not ordinal. They are separative objectives within the project.

** I will develop simple baselines model early-on for all the objectives.

1) Data Processing:

- Remove unnecessary symbols and stop words from text

2) Feature Engineering / Selection
    - Perform TF-IDF on the product description.
    - It seems that loads of information – e.g. quantity and weight – are buried within the product description.
    - Apply feature selection metrics like RFE and Select_KBest to assess features.

3) Exploratory Data Analysis
    - Do features like product type, source, and/or destination correlate w/ sale price?
    - Do product types correlate with seller's ratings or country source/destination?
    - What are common traits of the most lucrative sales?
    - Analyze price distributions.

4) What are the characteristics of the top performers?
    - Apply a regression – e.g. linear, decision tree, random forest, and XGBoost – to predict the selected metric that determines the

5) Identify outlier within the sales.
    - Did a product sell for a significant discount compared to similar products?
    - What are the traits of these abnormal transactions?
        – Are there similarities of how the products are described?
        – Are there certain sellers who are associated with the anomalous transactions?

6) Develop product price estimate model.
    - Create a train, validation, and test set. Use train to train the model. Use validation to optimize hyperparameters by applying RandomizedSearchCV. Evaluate the model using the test set.
    - Models that will be used – linear regression, Decision Tree Regress, Random Forest Regression, and XGBoost Regression.

7) Deploy "Morgan's Back Book" app. Potential buyer/seller can enter parameters and receive a price quote on the product.
    - User can input product type, description and other attributes, and the model will provide a price range