Investigating Enron Email Dataset

MSDS 422 Practical Machine Learning

Clark Morin, Madison Trimble, Trey Pezzetti, Julie Athanasiadis

March 10, 2024

PRESENTATION TOPICS

1. Research Problem
2. Research Objectives
3. Literature Review
4. Approach
5. CRISP-DM Phases
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
    - Deployment
6. Findings and Conclusions
7. Limitations, Recommendations, and Future Work
8. References

Research Problem

- In the wake of the Enron scandal, understanding the dynamics and characteristics of internal communications can provide insights into the operational behaviors that precede corporate crises.

- This project aims to investigate the Enron Email Dataset to uncover patterns, behaviors, and key markers that could serve as indicators of corporate health or misconduct.

Research Objectives

1. To perform a comprehensive exploratory data analysis (EDA) to understand the structure and nature of Enron's internal communication data.

2. To analyze sentiment and thematic changes over time to see if they correlate with known events in the timeline of Enron's collapse.

3. To identify central figures and the structure of the communication network within Enron using social network analysis.

4. To apply topic modeling to the corpus to identify prevalent topics and their evolution.

5. To develop a model for anomaly detection that can potentially flag unethical behavior patterns in corporate email communications

Literature Review

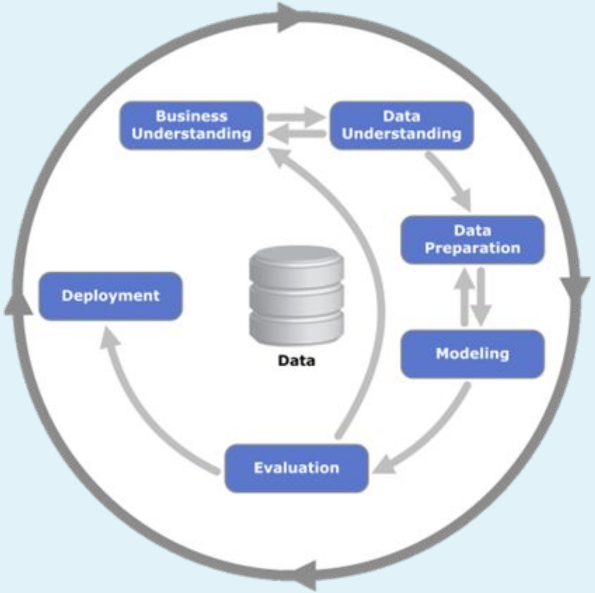
- ***Random Forests Machine Learning Technique for Email Spam Filtering* article by E.G. Dada and S.B. Joseph**
  - Content-based filtering approach that is used to identify certain keywords that may be frequently associated with phishing or spam within emails within the dataset.
  - Random Forests algorithm to achieve feature extraction and classification.
  - This model's performance was evaluated using 10-fold cross validation and achieved an accuracy of 99.92% and an error rate of 0.0296% (Dada and Joseph 2018).

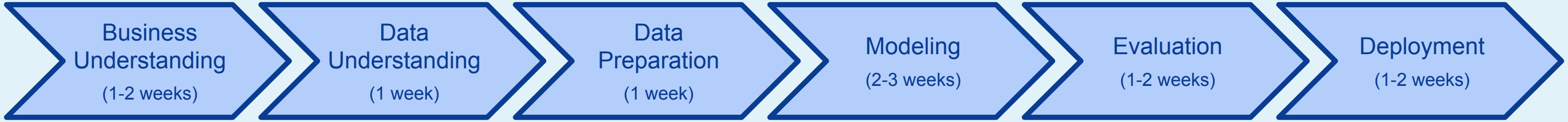
- **2016 study by Shrawan Kumar Trivedi**
  - Multiple machine learning tools were employed to determine which performed best in classifying emails based on the presence of certain features.
  - Email files were extracted from the database and transformed by removing HTML tags, stop words, and converting words to their base form, also known as lemmatization, to create a feature dictionary (Trivedi 2016).
  - Machine learning classifiers Bayesian, Naive Bayes, Support Vector, and Decision Tree. Each model's performance was measured by accuracy, false positive rate, and training time. Support Vector achieved the lowest false positive rate at 6.5% and the highest accuracy at 93.3% (Trivedi 2016).

Approach

Agile project practices with **CRISP-DM** approach

- ✔ Iterative Development
- ✔ Flexibility
- ✔ Ongoing Value Delivery





Business Understanding (1-2 weeks)

Data Understanding (1 week)

Data Preparation (1 week)

Modeling (2-3 weeks)

Evaluation (1-2 weeks)

Deployment (1-2 weeks)

CRISP-DM Project Phases: Cumulative Estimated Timeframe (~11 weeks)

Business Understanding

Cumulative Time: ~2 weeks

1-2 weeks

Business Objectives

- Suspected fraud amongst Enron employees
- Assist compliance officers in investigating any wrongdoing by exploring employee emails

Assess Situation

- Python was the primary programming language chosen for coding and executing our models.
    - Python offers extensive libraries, versatility, and readability necessary for coding algorithms and collaboration.
- Cloud-based coding environment enabled collaboration on code development in real-time and allowed for simultaneous access to the updated code.
- Google Colab provided access to GPU resources for faster model training and execution.
    - The default GPU for Colab is the NVIDIA Tesla K80 which provides 12GB of VRAM.
- Risks, assumptions, and constraints
- Costs and Benefits

Determine Mining Goals

- Machine learning and deep learning techniques to develop a categorical predictor that flags potential fraud in email communications.

Project Plan

- CRISP-DM Model
- Stakeholder collaboration

Data Understanding

Cumulative Time: ~3 weeks

1 week

Collect Initial Data

- Available per government for use in research

Describe Data

- 10,000 emails
- Datatypes come over as categorical, boolean, and numerical
  - (To, From, Subject, Body, Date)
- Body of Email has a large number of text on body of email

Explore Data

- Plotting Email Volume by Month Over Time
- Top Senders and Receivers of Emails
- Sentiment Analysis via TextBlob
- Subjectivity Score Analysis

Verify Data Quality & Batch Processing

- Reliable datasets and infrastructure
- Search for Incomplete/incorrect data
- Sampling bias if data is not large enough

Data Preparation

Cumulative Time: ~4 weeks

1
week

Select Data

- Evaluate sender and receiver information, body of email, and date

Clean Data Confirmation

- Extraction of Key Email Attributes:
  - The script defines a function to parse each email and extract essential attributes such as the "From" (sender), "To" (recipient), "Subject", "Date", and "Body". This step transforms email data into a structured format suitable for analysis

Construct Data

- Creation of the Fraud_Flag column:
  - This binary feature indicates whether an email potentially relates to financial fraud, determined by the presence of specific keywords related to financial fraud in the email's subject or body
- Calculation of email length:
  - determined by the number of characters in the email's body.
- Email Domain Extraction:
  - The script performs a transformation to extract the domain from each email address in the datasets.
- Sentiment Polarity Score:
  - Sentiment analysis is conducted on the email bodies, resulting in a sentiment polarity score for each email. This score is a continuous variable that measures the sentiment tone of the email, ranging from negative to positive.

Format Data

- Normalization of Email Length:
  - To standardize the email length feature for model training, the script applies normalization to this variable

Modeling

Cumulative Time: ~7 weeks

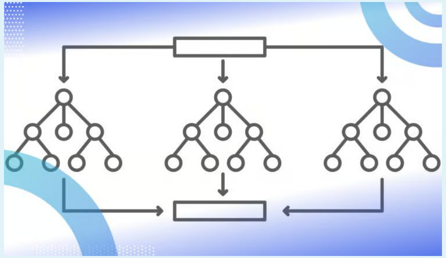
2-3 weeks

Select Modeling Technique

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- Support Vector Classifier

Generate Test Design

- Goodness of Fit
- Separate data into train and test data. Build model on train set and estimate it performance on test set



Build Model

- Scikit-Learn
- Tensor Flow
- Keras
- TextBlob

Assess Model

- Cross Validation, precision, recall, confusion matrix, F1 score, and ROC AUC.

Evaluation

Cumulative Time: ~9 weeks

1-2 weeks

Evaluate Results

- Make sure results and findings are clear and easy to understand
- Measure model results against business objectives
- Rank Model performances
- Highlight any unique findings

Approved Models and Determine Next Steps

- Determine models for deployment
- Decide to move forward to deployment or go back and refine or replace some your models.

Deployment

Cumulative Time: ~11 weeks

1-2 weeks

Deployment Planning

- Summarize models and findings
- Create plan to disseminate information to stakeholders

Plan Monitoring and Maintenance

- Create plan to monitor and maintain model. Revisist with new discoveries.
- Determine expiration of models if applicable

Final Report and Deploy

- Summary of deliverables and results
- Presentations to stakeholders

Final Review

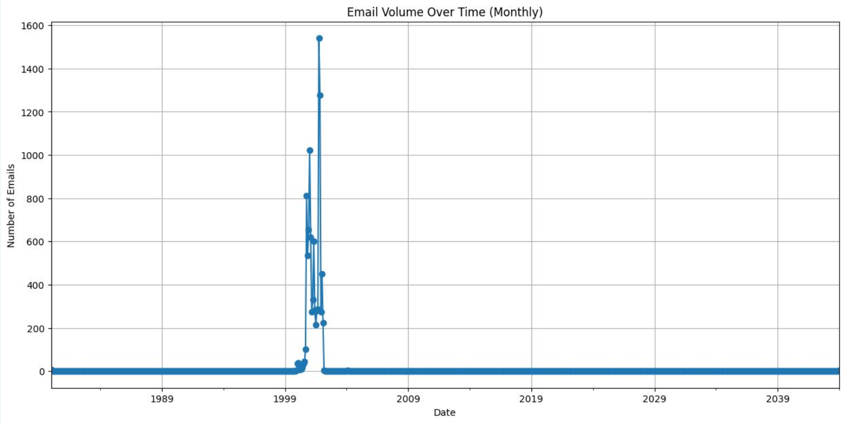
- Store Project Documentation
- Limitations and Suggestions for future work

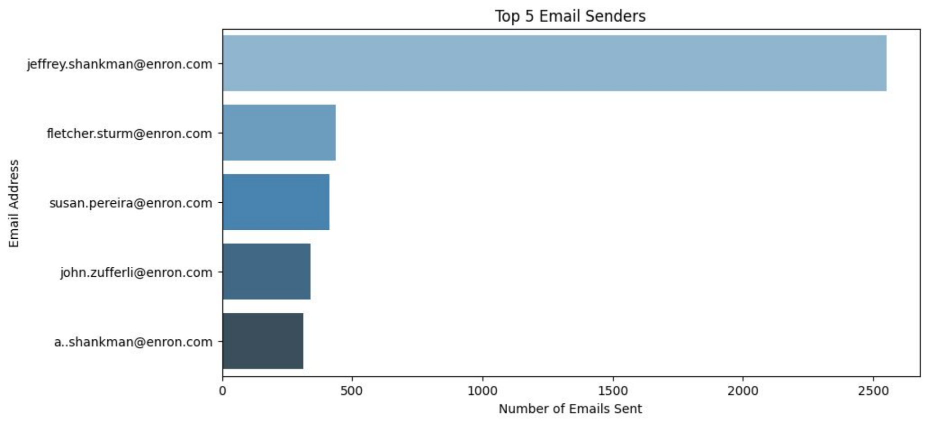
Findings and Conclusions

EDA

- **Plotting Email Volume by Month Over Time**
  - Found no relevant patterns more research needed

- **Top Senders and Receivers of Emails**
  - Found top 5 senders and receivers but more research needed to determine significance.

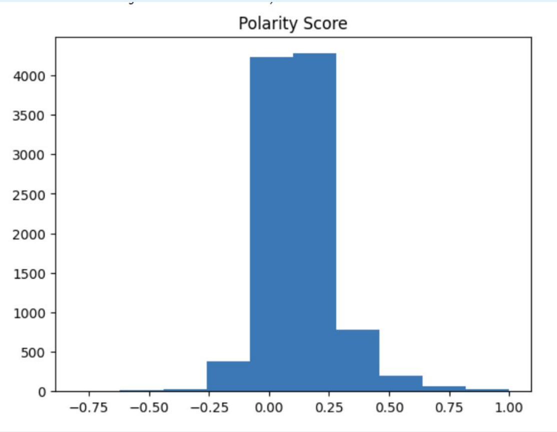


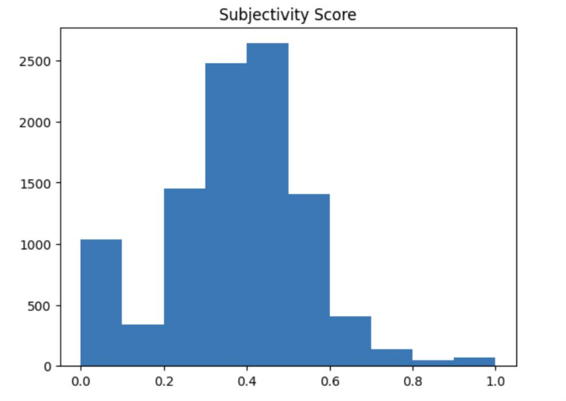


Findings and Conclusions Continued

EDA

- **Sentiment Analysis via TextBlob**
  - most emails have polarity scores between -0.25 and 0.5, suggesting a predominantly neutral sentiment. However, emails with scores below -0.5, indicating negative sentiment, could be critical for identifying communications that might negatively impact internal culture
- **Subjectivity Score Analysis**
  - most emails fall between 0.2 and 0.8 in subjectivity, pointing to a moderate level of subjectivity in communications. Notably, around 1000 emails were found to be completely objective. Emails with subjectivity scores above 0.8, marking them as highly subjective, present an interesting subset for further review due to their potential implications for organizational culture





Findings and Conclusions Continued

Model Results

- Accuracy: Each of the models exhibited a very high level of accuracy in their performance across the different cross-validation folds
  - linear regression mode 99.6%
  - random forest 99.8%
  - gradient boosting 99.9%.
  - support vector classifier model 99.7%

- Each model has a standard deviation of 0.0 meaning that they all exhibited minimal variability.

- Assessment Measurements
  - confusion matrix using the Random Forest Classifier returned 6 true positives, 0 false positives, 1987 true negatives, and 7 false negatives
  - precision ratio of 1.0 indicates that all predicted positive cases were indeed positive.
  - Recall of 0.4615 shows that the model was able to determine 46.15% of actual fraudulent emails.
  - F1 score of 0.6316 indicates a relatively balanced measure of precision and recall.
  - ROC AUC score of 0.9966 suggests that the model has high discriminatory power to distinguish between positive (fraudulent) and negative (not fraudulent) classes.

- **The scores listed above indicate that the models can effectively generalize from the dataset. The decision regarding which model to deploy could come down to the combined value of accuracy, interpretability, and computational efficiency.**

Limitations, Recommendations, and Future Work

Limitations

- Evidence in this research limited to emails only. Does not include telephone calls, in-person discussions, or physical documentations.
- A lot of “noise” in the emails and a lot of neutral emails. Small number of true positive in confusion matrix. Models need to be refined to capture important information.

Recommendation Actions

- Share findings with vital stakeholders
- Use findings as evidence for discipline and legal action
- Conduct interviews with employees with flagged emails
- Place safeguards into place to prevent happening again

Future Work

- Apply model to other cases of fraud detection
- Use deep learning methods to strengthen model and approach
- Third party datasets that can add value to the existing analysis

References

Data Source:
https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22

Data Source:
https://ieeexplore.ieee.org/abstract/document/9489015

Data Source:
https://www.researchgate.net/profile/Emmanuel-Dada/publication/327720311_Random_Forests_Machine_Learning_Technique_for_Email_Spam_Filtering/links/5ba0b0c0299bf13e6038ecc1/Random-Forests-Machine-Learning-Technique-for-Email-Spam-Filtering.pdf

CRISP-DM Model Visual:
https://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf

CRISP-DM Model Guide:
https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide

Meet the Team

Clark Morin

Julie Athanasiadis

Madison Trimble

Trey Pezzetti