**Project Plan and Team Assignment: Email Anomaly Detection Using ML**

**Phase 1: Literature Review & Learning (Week 1–2)**

**Objective**: Understand existing research in email-based phishing/spam detection and identify key challenges.

Responsible: Shiyun Liu，JuanDu

-Read and summarize at least 5 key research papers on phishing detection.

-Identify relevant features, ML methods, and evaluation strategies.

-Maintain a “Literature Review” section on GitHub Wiki.

**Phase 2: Data Understanding (Week 2)**

**Objective**: Explore and interpret the dataset structure and main attributes.

Responsible: Alpar Arman，Wenbo Wu

-Investigate data columns (subject, sender, body, headers, etc.).

-Use visualization tools to analyze distributions and patterns.

-Report data quality issues (missing values, imbalance, etc.).

**Phase 3: Preprocessing and Feature Engineering (Week 3–4)**

**Objective**: Clean and transform email content into meaningful input features.

Responsible: Alpar Arman，Onkar Kadam

-Perform text cleaning (HTML removal, tokenization, stop word filtering).

-Engineer features (BoW, TF-IDF, hyperlink count, stylistic stats, etc.).

-Output processed dataset and upload to GitHub.

**Phase 4: Model Training with Traditional ML (Week 4–5)**

**Objective**: Train baseline models using engineered features.

**Responsible:** AlparArman**，**JuanDu**，**Shiyunliu

-Train Naive Bayes, Logistic Regression, and Random Forest.

-Use cross-validation and test set evaluation (AUC, confusion matrix, balanced accuracy).

-Report and compare model performance.

**Phase 5: Model Optimization and Validation (Week 5–6)**

**Objective**: Improve performance and address data imbalance.

Responsible:  WenboWu ，Oncar Kadam

-Apply SMOTE/under sampling for balancing classes.

-Tune hyperparameters and apply regularization techniques.

-Validate using hold-out sets and analyze results.

**Phase 6: Embedding-based Modeling (Week 6–7)**

**Objective**: Use embeddings (Word2Vec/BERT) to enhance semantic feature representation.

**Responsible: Alpar Arman，Onkar Kadam，Wenbo Wu**

-Select and implement word embeddings.

-Replace BoW features and retrain ML models.

-Compare embedding-based results with traditional approaches.

**Documentation and Presentation Tasks**

| **Task** | **Responsible** | **Notes** |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| GitHub contributions | All members | Commit code, open issues, document progress |  |  |  |  |
| Technical report writing | Alpar Arman，  Wenbo Wu | Literature, methodology, results |  |  |  |  |
| Final presentation slides | Onkar Kadam Shiyun Liu JuanDu | Slide design，visual summary |  |  |  |  |
| presentation | All members |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Weekly Timeline Overview**

| **Week** | **Activity** | **Deliverables** |
| --- | --- | --- |
| 1 | Literature review + data load | Research summary, dataset overview |
| 2 | EDA + basic preprocessing | EDA report, preprocessing scripts |
| 3 | Feature engineering + baseline model | Feature set, initial models |
| 4 | Model optimization | Performance logs, tuning reports |
| 5 | Embedding modeling | Embedding code, evaluation results |
| 6 | Report and slides drafting | First draft of report and presentation |
| 7 | Final edits + mock presentation | Final report + slide deck |