**Project Report: AI Compliance Email Classifier with SHAP Explainability**

**Author**: Aditya  
**Date**: 4th July 2025

**Objective**

The main goal of this project was to build a small AI system that can read internal company emails and decide if the content is:

* Compliant
* Needs Review
* Escalation Required

This is a real use case in finance or banking domain where internal communications must follow policy and legal compliance. I tried to simulate how AI can support this process.

I created this project after applying for an AI Analyst role to show my genuine interest in the company, in responsible AI, and how I think about real-world problems.

**What I Did**

**1. Created Sample Email Data**

I manually wrote and labeled 50+ pretend internal emails based on realistic financial and compliance tone. Each one was marked as Compliant, Needs Review, or Escalate.

**2. Converted Text to Numeric Features**

I used a method called **TF-IDF** to turn email text into numbers. This is needed for machine learning models to understand the text.

**3. Trained a Model**

I used a **Random Forest Classifier** (from scikit-learn) to learn from the labeled data. After training, the model gave around **63% accuracy** on test data which was decent for a small dataset.

**4. Explained Model Decisions with SHAP**

After training, I wanted to **understand why** the model gave a prediction. I used **SHAP**, a tool that explains which words influenced the prediction the most.

SHAP shows a **bar chart** of top features (words) that helped push the model toward a certain decision (e.g., Escalate). For example, if a word like “skipped” appears in an email, SHAP may show it as pushing the prediction toward Escalate.

**Why the SHAP Chart Showed +0.00 for All Bars**

In one of the test emails, the chart showed +0.00 for all features. At first, I didn’t understand why but later I learned this happens when:

* The model is not strongly confident in one direction
* Or the words are neutral, not risky or clearly safe
* Or the SHAP impact values are very small, like +0.0001 (so it rounds to 0.00)

This is normal behavior. I then changed the test sample (i = 0 to i = 2) and got a better chart with real impact numbers like +0.2, -0.1, etc.

It helped me understand that **not every prediction is strong**, and **transparency is just as important as accuracy**.

**Tools Used**

* Python
* Jupyter Notebook
* scikit-learn for modeling
* SHAP for model explanation
* TF-IDF for text processing

**Conclusion**

This project was my way to show initiative while waiting to hear back about the AI Analyst role. I chose a real-world use case in the finance domain, designed a working model, and made sure it was **explainable and responsible**.

This matches the job description in key areas:

* Model development and testing
* Explaining AI outputs to non-technical teams
* Promoting responsible and interpretable AI
* Applying AI to pain points like compliance

I enjoyed doing this, and I would love to work on larger problems like this inside your team.

Let me know if you'd like a walk-through of the notebook, or if I can answer anything about this project. Thanks for reviewing!

—  
**Aditya**