Assessment Project Full Report

SEC-301: Security Challenges in Modern AI Systems

Team Member [5 Points]:

1. [Your Full Name] – [Your Nickname] – [Your Email Address]
2. [Your Full Name] – [Your Nickname] – [Your Email Address]

AI System Description [10 Points]:

Write a paragraph describing the AI system you are going to develop for this assessment. This paragraph should serve as an executive summary, providing a high-level overview of the system's functionality and architecture. It must address the following questions:

- What are the key functionalities or objectives of the AI system?
- What are the inputs and outputs?
- How could an attacker exploit the system or derive benefits from it?

A contextual diagram may be included at the end of the paragraph to illustrate the system's overall architecture.

A List of Datasets [5 Points]:

Provide a numbered list that shows all datasets that will be used to train and develop the AI system. For example:

1. [Dataset A Title] – [Link to download the dataset]
2. [Dataset B Title] – [Link to download the dataset]
3. [Dataset C Title] – [Link to download the dataset]

Development of the AI System [45 Points]:

Write a section of multiple paragraphs to explain and describe how you develop the AI system. It must explain, with your thinking process, how to deal with the requirements and process data. You must provide screenshots of your source code/your source code fragments and explain each part comprehensively.

Example

```
# Impute missing numeric values with median
numeric_cols = ["numeric_feature", "supp_feature1", "supp_feature2"]
df_raw[numeric_cols] = df_raw[numeric_cols].apply(
    lambda col: col.fillna(col.median())
)

# Replace malformed text entries
df_raw["text_field"] = df_raw["text_field"].replace(r"\s+", " ",
regex=True).str.strip()
# Clip extreme outliers (99th percentile)
for col in numeric_cols:
    upper = df_raw[col].quantile(0.99)
    df_raw[col] = df_raw[col].clip(upper=upper)
```

After we successfully download the datasets from the Internet, we conduct an exploratory data analysis task to understand and identify data quality issues. We found that there are missing values, malformed texts, and outliers, which are considered as data quality issues in this project. We implement a Python source code to get rid of each issue one-by-one. Starting with the imputation of the missing value, we found that the median is the best way to impute the dataset as it will not significantly affect the important insight. Then, we use regular expression to replace malformed texts with a white space. Lastly, we identified outliers that extremely exceed 99^{th} percentile and then clip them.

Output Screenshot of the AI system [30 Points]:

Provide a collection of screenshots from the working AI system and write a multi-paragraph section that explains each output, indicating whether it matches the predefined objectives and functionalities. This section should be detailed enough to convince readers that the system operates as expected and has been properly developed.

Security Threat Analysis Techniques [20 Points]:

Write a section of paragraphs that discusses the security threat analysis techniques covered in the lectures. Then compare those techniques with respect to the target AI system and identify which technique(s) are most suitable. Provide a justification to support your selection of the security threat analysis technique(s).

Security Threat Analysis Procedure [50 Points]:

Write a section of paragraphs or multiple sub-sections to explain your security threat analysis in detail. Provide justification for every decision made during the analysis, and include intermediate results that reveal your thinking process. A section that presents only the final list of security threats will receive a significantly reduced score.

Example

We use the STRIDE model together with data-flow analysis. First, we identify the components within our AI system. The main component is the AI model that we train on the datasets. Another important component is the user interface, where end users can send prompts to the model (via the command-line program) and receive the response back in text form. Next, we draw security boundaries between internal/trusted and external/risky components, because we want to see how each component can be threatened during its operation. ...so on.

Output of the Security Threat Analysis [30 Points]:

Write a section of paragraphs to report all the outputs of the security-threat analysis performed in the previous section. List every identified threat and analyze them in order to select a subset that you will exercise in this assessment project. Dedicate a paragraph to emphasize the selected threat(s) and provide justification for your choice.

Example

From the security-threat analysis reported above, we identified five potential threats to our AI system:

1. **Data poisoning** – This threat is plausible because the model is trained on public datasets whose provenance cannot be fully verified. Although we conduct data analysis and cleaning to detect outliers and suspicious records, the process cannot guarantee that all poisoned samples are eliminated.
2. ...

At the end, our team selects the data-poisoning attack as the primary focus because it is both likely and difficult to mitigate in the AI system we develop. Data-poisoning attacks are common, can be introduced during the data-collection phase, and often evade standard cleaning procedures, making them an especially relevant and challenging threat to address.

Attack Scenarios and Procedure [50 Points]:

Write a section of paragraphs that explains the attack procedure you will use to target your AI system. Provide as many details as possible so readers can understand and reproduce the attack. The explanation must include the preparation (e.g., reconnaissance), execution, and exploitation stages. Carefully describe how an attacker could exploit the selected threats; omitting key steps often makes the attack scenario unclear for readers. You may provide images or samples of attack code.

Example

In the preparation stage, we will create a clean dataset of pseudonymised COVID-19 patients and their symptoms. Then we define our attack goal as causing the system to misclassify certain rare cases and to output patient information ...

Outcomes and Consequences of Attack [40 Points]:

Write a section of paragraphs that reports the results and outcomes of the successful attack. Include a collection of screenshots that demonstrate the attack attempts and confirm that the attack was executed correctly. In addition, analyze the consequences for the AI system and the benefits gained by the attacker after successful exploitation.

Survey on the Mitigation of the Selected AI-Specific Attack [50 Points]:

Write a literature review section that explores recent research papers (2020 or later) on protecting AI systems and mitigating the selected attacks. Survey sources such as Google Scholar, arXiv, IEEE Xplore, or ScienceDirect, and include at least five papers to ensure broad coverage of the field. Conclude with a dedicated paragraph that summarizes the main findings of the survey.

Selected Attack Mitigation Procedure [40 Points]:

In this section, you are required to describe and justify the mitigation procedure(s) applied to defend the AI system against the selected attack identified in the previous sections. The purpose of this section is to demonstrate your understanding of AI-specific security defenses, your ability to apply appropriate countermeasures, and your capacity to reason about their effectiveness and limitations.

Your discussion should focus on primary attack(s) selected earlier (e.g., data poisoning, model inversion, prompt injection, evasion attacks, membership inference) and must be aligned with the threat analysis and attack scenarios already presented.

Output of the AI System After Implementing Security Measures [20 Points]:

In this section, you are required to provide visual and explanatory evidence demonstrating that the AI system is operating after the selected security mitigation procedures have been successfully implemented. The purpose of this section is to confirm that the system remains functional while exhibiting improved resistance to the selected attack.

You must present screenshots taken from the protected version of the AI system and explain how these outputs reflect the applied security measures and their effectiveness.