**Risk Report**

**CDS 492- 002**

**Spring 2024**

**Identified Risks:**

1. Data Quality Issues:

* Incomplete or inaccurate data in the job posting dataset could affect model training and performance.
* Imbalanced data between genuine and fake job postings may bias the model towards the majority class.

1. Model Performance Issues:

* The model may fail to accurately detect fake job postings due to complex patterns or evolving techniques used by malicious actors.
* Overfitting or underfitting of the model to the training data could result in poor generalization to unseen data.

1. Resource Constraints:

* Insufficient time and resources for data analysis, model development, and evaluation could delay project timelines.

**Risk Mitigation Strategies:**

1. Data Quality Assurance:

* Implement data preprocessing techniques such as handling missing values, removing duplicates, and standardizing data formats to improve data quality.
* Conduct thorough exploratory data analysis to identify and address data inconsistencies.
* Establish data quality checks and validation procedures at each stage of the project lifecycle.
* Address the imbalance between genuine and fake job postings by employing techniques such as resampling or using ensemble methods, such as Random Forest and gradient boosting.

1. Model Performance Optimization:

* Experiment with different machine learning algorithms, feature engineering techniques, and hyperparameter tuning to improve model performance.
* Regularly monitor model performance metrics and retrain the model with updated data to adapt to changing patterns.

1. Resource Management:

* Prioritize tasks based on their criticality and allocate resources efficiently to meet project milestones.
* Consult the professor and peers to find the best approaches to optimize time and resources as needed.

**Bias and Fairness Considerations:**

* Evaluate the dataset for potential biases related to demographic factors, job categories, or geographical locations.
* The potential bias arises from the significant class imbalance in the dataset, with only 800 instances of fake job postings out of 18,000 data points, potentially leading the model to prioritize the majority class (real job postings) and overlook the minority class (fake job postings). This can be addressed by utilizing machine learning techniques like Random Forest and Gradient Boosting, which are adept at handling such data, to improve the model's ability to accurately classify both majority and minority class instances in fake job posting detection.
* Document and disclose any biases or limitations in the model to stakeholders to promote transparency and accountability.

**Transparency Measures:**

* Document the entire model development process, including data collection, preprocessing, feature selection, and model training.
* Provide clear explanations of model predictions and decision-making criteria to stakeholders.
* Encourage open communication and feedback mechanisms to address concerns and improve model transparency over time.