sofvie

Module-7 Mid-Term Presentation

Capstone (ALY6980)

College:
College of Professional Studies, Master of Professional Studies in Analytics

Current Academic Term:
5th Quarter (Spring 2023)

Guided By:
Prof. Jay Qi

Submitted By:
1. Parth Savaliya (002982302)
2. Dhimahi Patel (002985259)
3. Parva Patel (002195186)
4. Pratikkumar Malaviya (002963548)

Group No: 4

Date of Submission: 21st May'2023

Northeastern University

AGENDA

DOMAIN

- Introduction
- Project Description
- EDAs

1

DATA

- Data Understanding
- Business Questions

2

MODELING

- Data preparation for Models
- Techniques
- Model and Evaluation

3

SUMMARY

- Conclusion
- Future Work

4

Northeastern University

INTRODUCTION

- The given dataset consists of observations related to **hazard incidents in the domain of safety management**. It includes various columns such as 'Hazard_type', 'Hazard_identification', 'Completion_Time_Bucket', 'potential_risk', and more.

- The dataset, comprising approximately **14,000 records**, which are divided into further **24 columns,** and it was generated by combining different tables from Sponsor Sofvie.

- The **target variable** of interest in this dataset is **'Completion_time_bucket**', which represents the time taken to complete the hazard incident.

- This dataset **provides valuable information for analysing hazard incidents**, identifying potential risks, and evaluating the effectiveness of immediate and recommended actions taken.

- The aim is to explore patterns and factors affecting completion time to enhance safety management practices.

Northeastern University

PROJECT DESCRIPTION

- The project focuses on incident and hazard management using machine learning techniques to predict and prevent incidents, as well as efficiently address them when they occur.
- It involves analysing incident **trends over time**, predicting the likelihood of incidents based on historical data and employee/supervisor details, and identifying patterns in the data to provide recommendations for reducing risks and incidents.
- The project also aims to classify different types of hazards, determine appropriate actions, and establish timeframes for completion.
- The ultimate goal is to enhance safety for individuals and communities by ensuring that events and **hazards are proactively managed and resolved in a timely manner.**

DATA UNDERSTANDING

Target Variable

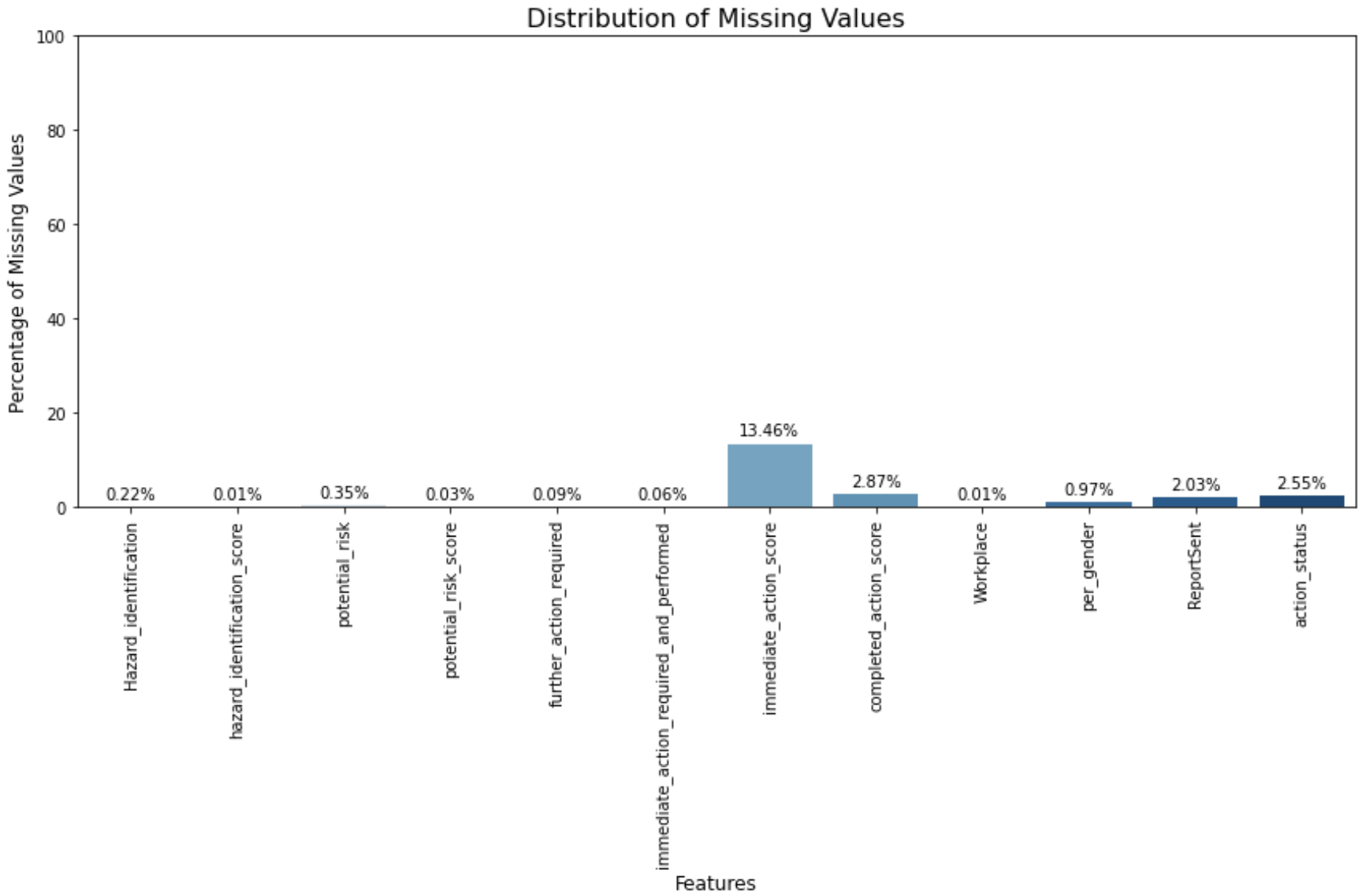
The target variable for analysis in this project is the **completion_time_bucket**. This variable represents the **time taken to address and resolve each hazard incident**. Understanding the factors influencing the completion time can help identify areas for improvement in incident management processes.

Dataset Source

The hazard data was collected from **various sources, including tables** from Sponsor Sofvie. These sources provide **comprehensive information** on incidents and hazards, ensuring the dataset's relevance to hazard incident management.

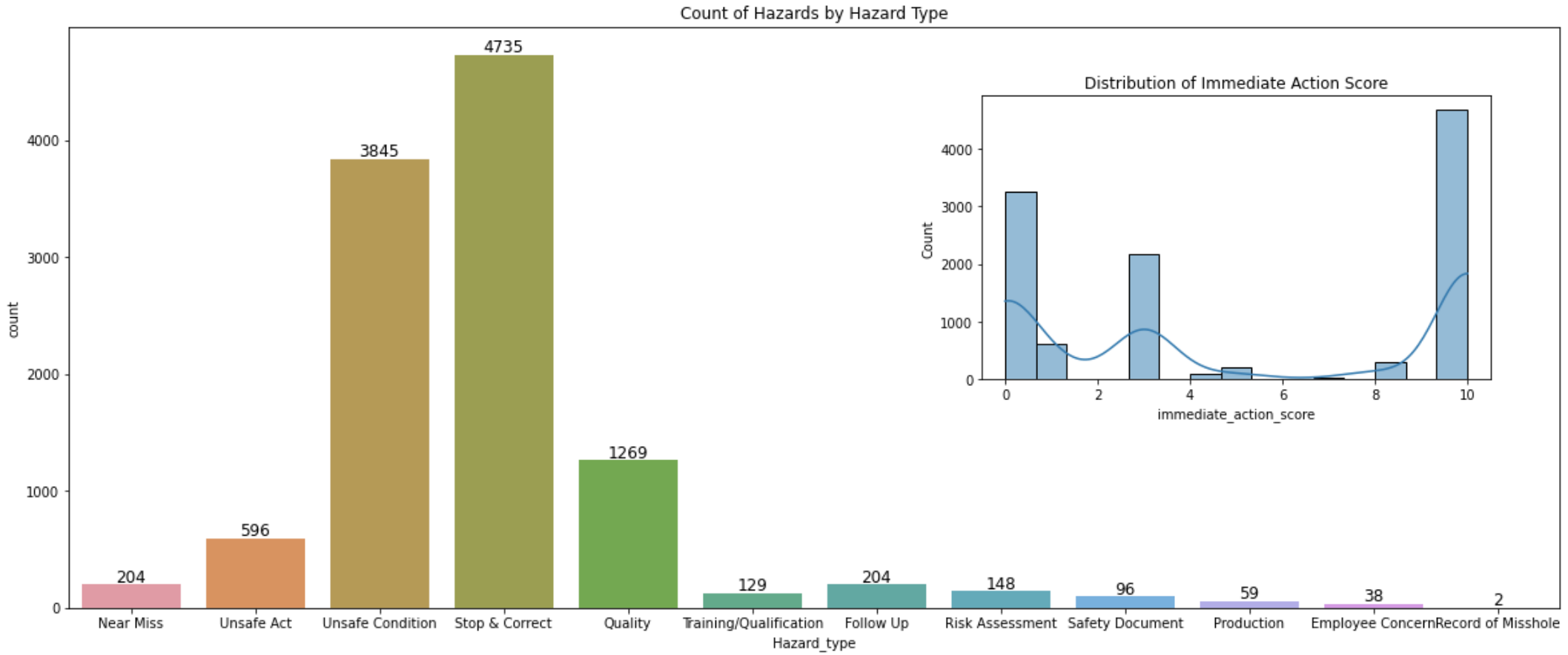
Northeastern University

EXPLORATORY DATA ANALYSIS



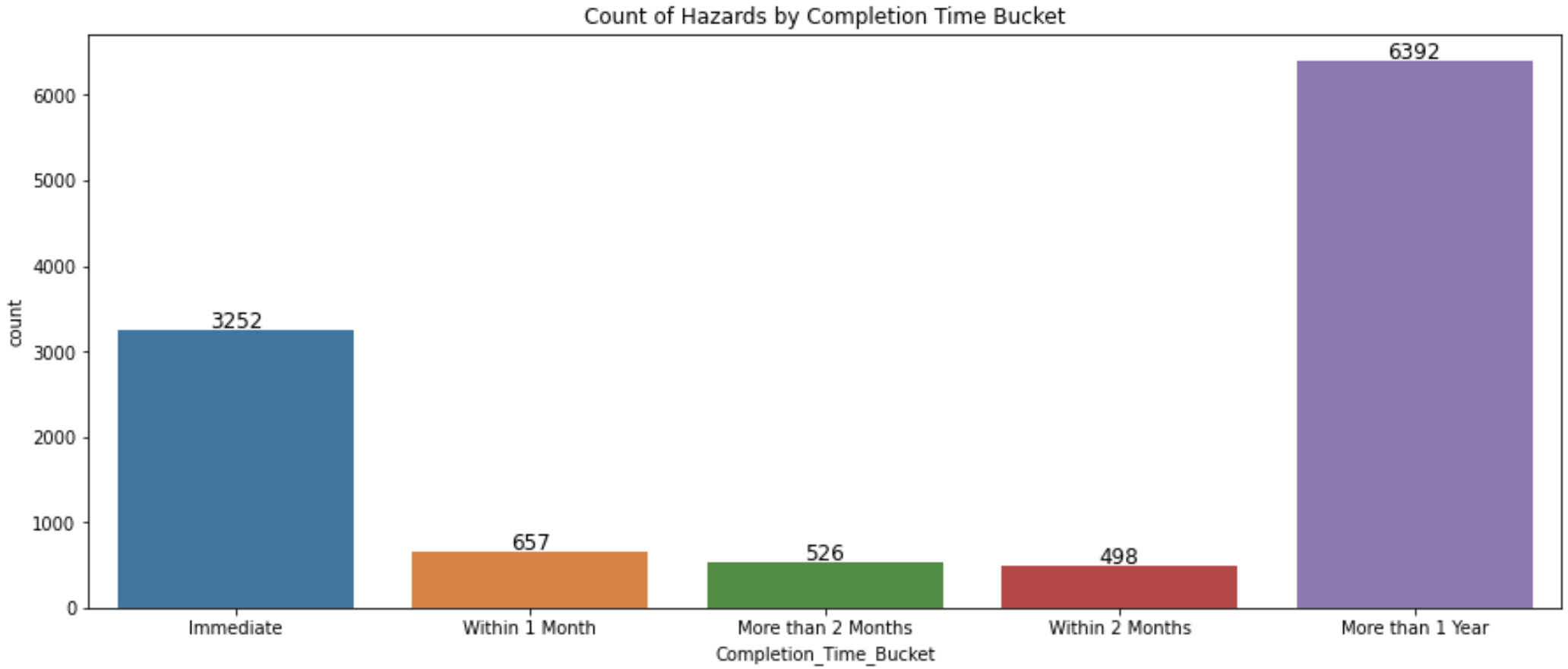
Northeastern University

EXPLORATORY DATA ANALYSIS



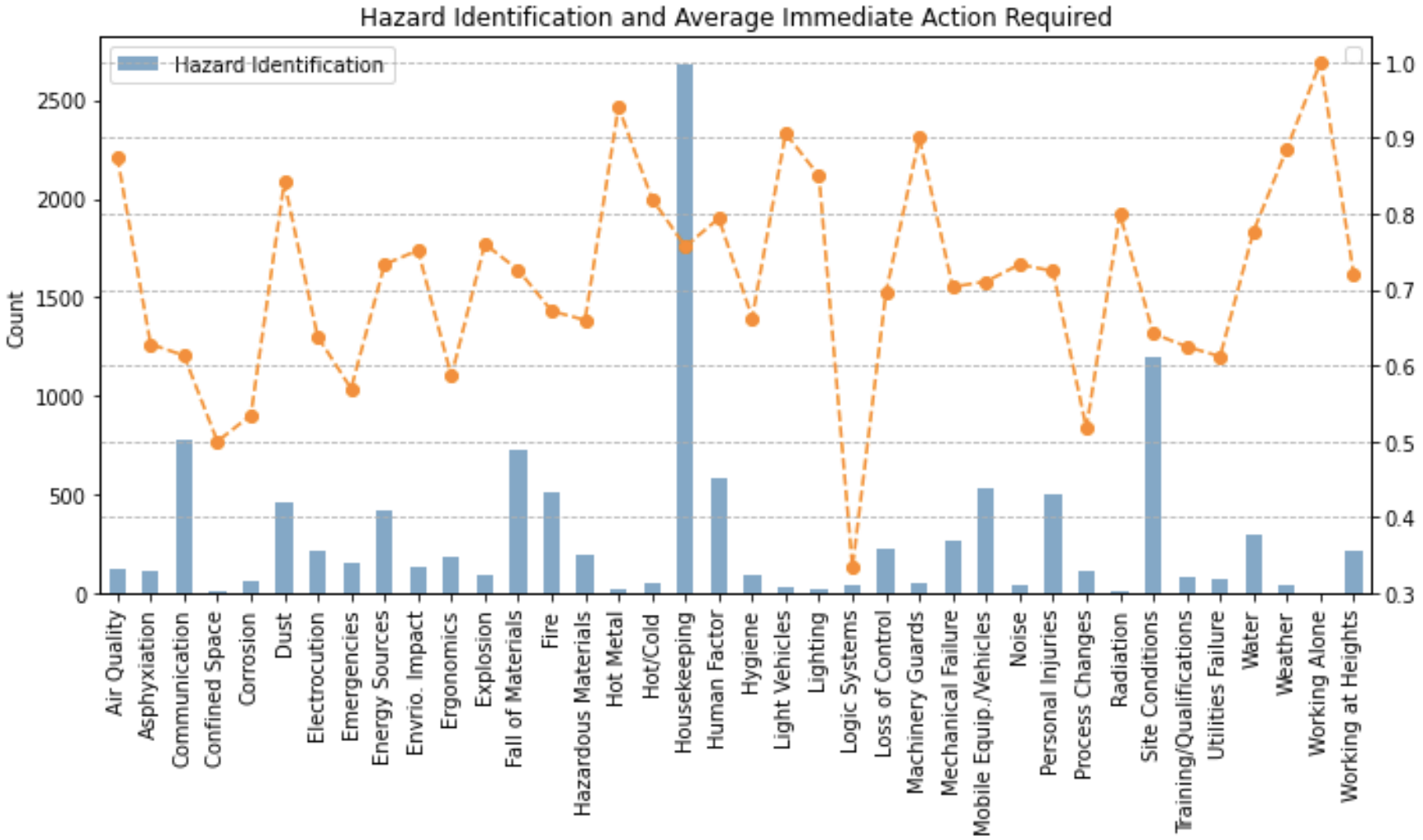
Northeastern University

EXPLORATORY DATA ANALYSIS



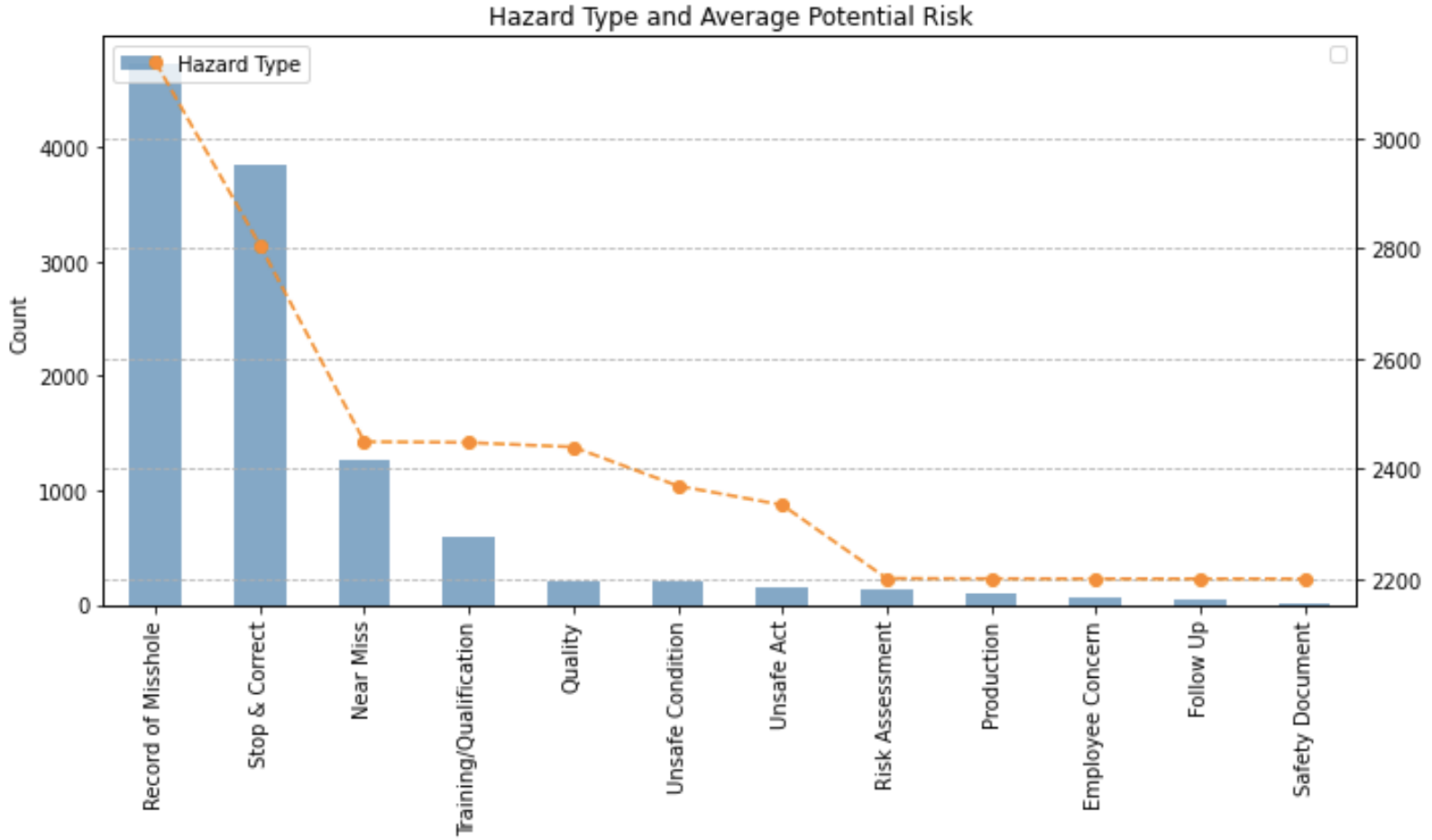
Northeastern University

EXPLORATORY DATA ANALYSIS



Northeastern University

EXPLORATORY DATA ANALYSIS



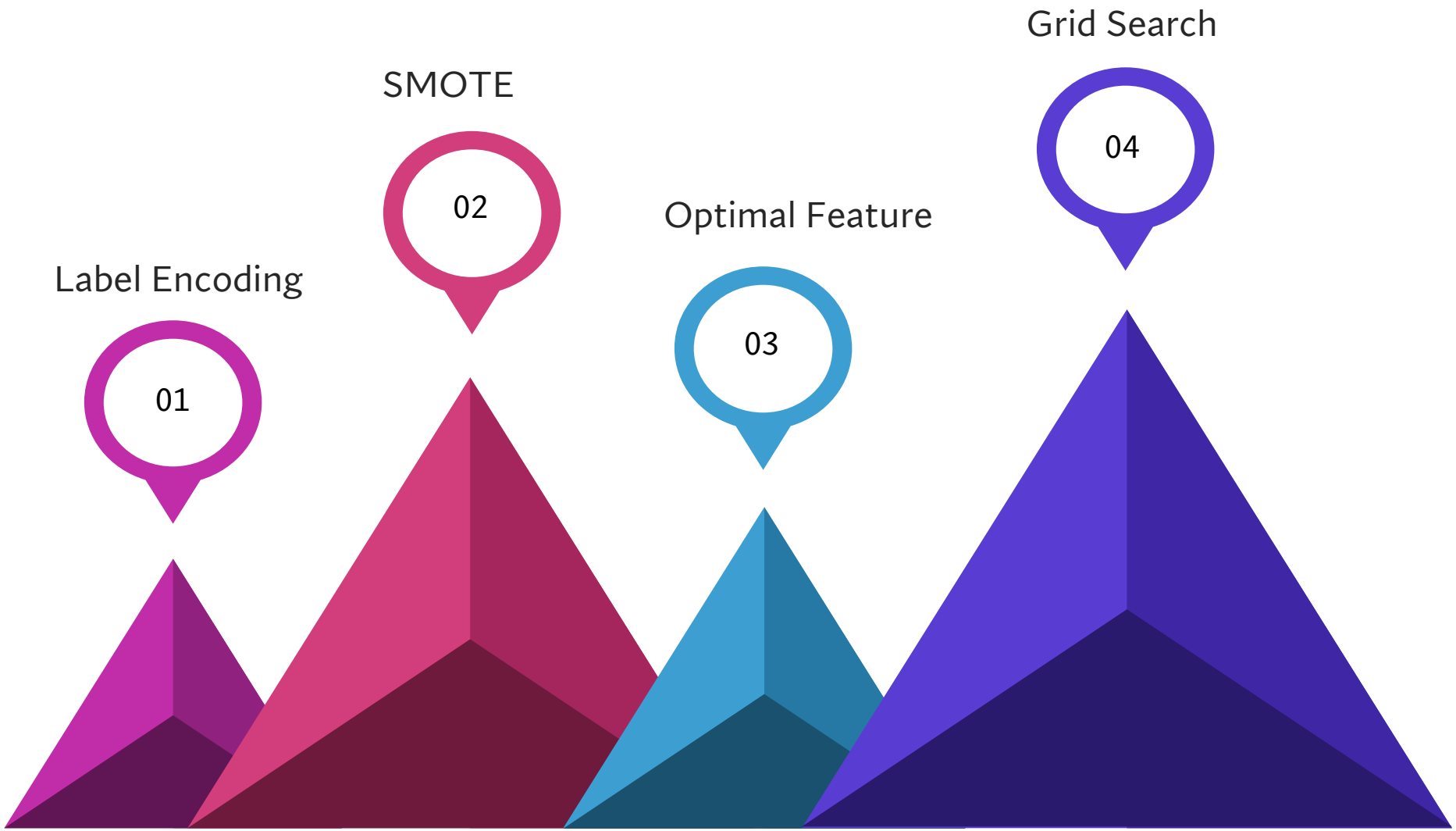
Northeastern University

BUSINESS QUESTION

Using the Trained Model, how company can predict the resolution time of the hazard which would help them to prioritize their work?

Northeastern University

POWERFUL TECHNIQUES



Northeastern University

DATA PREPARATION FOR MACHINE LEARNING

Label Encoding

In order to prepare the dataset for the ML model, we applied label encoding to categorical variables such as

- 'Hazard_identification',
- 'immediate_action_taken',
- recommended_action',
- 'Workplace',
- 'Completion_Time_Bucket',
- 'Hazard_type',
- 'Supervisor_Name'.

SMOTE

To address class imbalance in the **'Completion_Time_Bucket'** variable (target variable), we used the SMOTE (Synthetic Minority Over-sampling Technique) algorithm to generate synthetic samples of the minority class, ensuring a balanced representation of different completion time buckets and improving the model's performance.

Northeastern University

DATA PREPARATION FOR MACHINE LEARNING

Recursive Feature Elimination with Cross Validation

In order to determine the **optimal number of feature**s for the ML model, **we applied** Recursive Feature Elimination with Cross-Validation **(RFE-CV**), which **systematically selects features based on their importance and performance,** ultimately improving the model's predictive ability and reducing dimensionality.

Grid Search

To find the **best combination of hyperparameters** for the ML model, we utilized **Grid Search**, which exhaustively **searches through a specified parameter grid** and evaluates the model's performance for each combination, allowing us to identify the optimal hyperparameters that maximize the model's accuracy or other desired metrics.

MACHINE LEARNING MODELS & EVALUATION

Model	Selected Features	Class	Accuracy	Precision	Recall	F1-Score	Execution Time
XGB Classifier (Imbalanced)	3	immediate	0.82	0.67	0.89	0.76	10.41 sec
		more than 1 year		0.98	0.96	0.97	
		more than 2 months		0.27	0.15	0.19	
		within 1 month		0.22	0.08	0.12	
		within 2 months		0.33	0.12	0.17	
XGB Classifier (Balanced)	15	immediate	0.81	0.71	0.83	0.77	28.39 sec
		more than 1 year		0.99	0.97	0.98	
		more than 2 months		0.20	0.17	0.18	
		within 1 month		0.27	0.19	0.23	
		within 2 months		0.25	0.17	0.21	

Northeastern University

MACHINE LEARNING MODELS & EVALUATION

Model	Selected Features	Class	Accuracy	Precision	Recall	F1-Score	Execution Time
Random Forest Classifier (Imbalanced)	15	immediate	0.83	0.66	0.94	0.78	19.17 sec
		more than 1 year		0.99	0.97	0.98	
		more than 2 months		0.19	0.08	0.12	
		within 1 month		0.24	0.06	0.09	
		within 2 months		0.48	0.08	0.14	
Random Forest Classifier (Balanced)	13	immediate	0.81	0.67	0.92	0.77	24.58 sec
		more than 1 year		1.00	0.96	0.98	
		more than 2 months		0.21	0.13	0.16	
		within 1 month		0.23	0.08	0.12	
		within 2 months		0.37	0.10	0.16	

Northeastern University

CONCLUSION

- Since there is only a **slight difference** that can be noticed, we identified with the last three classes, which are **more than two months, within one month, and within two months**, for the Random Forest and XG-Boost models, which were created with two types of data, balanced and imbalanced.

- Both machine learning models performed well when they were compared, but the XG-Boost model has better result matrices when compared to the other one.

Recommendation : *We would advise our sponsor to train the **XG-Boost model** to forecast the resolution time based on the business question.*

Future Work : Now, let's move on to our next strategy,

- We would like work on the same dataset by dividing in before and after COVID-19 situation.
- Will try to analyze COVID-19 data and find appropriate patterns which discloses more insights.
- Moreover, we will try optimize prepared models with other suitable techniques.

Thank you!

Northeastern University