UNIVERSITY of HOUSTON

CNST 6308 – Data Analysis in Construction Management

Predicting Construction Accident Severity Using Classification Techniques

Submitted by

SURYA PUNNA – 2348592

Under the guidance of

Dr. Lu Gao, Ph.D.

University of Houston -Main Campus

Abstract

The report “**Predicting Construction Accident Severity and Identifying Contributing Factors Using Machine Learning**” is one the most important research in enhancing safety measurement and reducing risks on construction sites. This approach uses machine learning to assess the severity of construction accidents based on important risk factors encompassing environment, task, humans and projects. Through the Frequency Encoding and One-Hot Encoding, the study is able to handle both nominal and ordinal data, thus has consideration on the important factor such as the types of event, human factors that may influence the event, and the type of project that is susceptible to event.

Random Forest Classifier and other professional models are applied to foresee the possibility of fatal or no fatal outcome of the accident. These techniques aid in determining which of these variables—namely the particular task in question, the nature of the injury, and environmental aspects—positively correlate with and cause severe accidents. A detailed approach is also established as part of the work by outlining a data-driven approach for risk forecasting and accident elimination.

The key hyperparameters which include the number of trees implemented in Random Forest, the parameters for accuracy, precision and recall guarantee the models’ optimum for real world use. Apart from the estimation of future accident severity, this predictive system can also give construction firms and safety officers insights on trends in the types and causes of injuries and their environments. In the long run, such an approach is useful for safety management since it can help to prioritize intended precautions more accurately and coordinate resources brought into using those precautions more efficiently.

Carrying out an exploration of actual accident information, the project shows the key factors influencing the severity of an injury. Informed decision making of construction site stakeholders reduces accident incidences, enhances saving of lives, and general construction site safety management.

Table of Contents

1. Introduction

1.1 Problem Statement

Construction sites are the main areas of focus that must be safe to enhance the safety of the workforce and enhance project outcomes. Nevertheless, construction accidents remain a problem that causes serious harm or even death at construction sites. Even with current protective initiatives, forecasting the extent of mishaps remains difficult because of the variety of hazards and tasks that are involved such as environmental factors, human error, and the specific nature of the tasks involved in an enterprise. The existing methods of evaluation and determination of the severity of an accident are mainly based on quantitative data and qualitative estimates, which give little or no possibility of anticipating accidents or high-risk causes in advance.

This project aims to create a system that can classify construction accidents based on severity whether fatal or non-fatal, using the factors in t event type, task assigned, human factors, and environment. This work is to offer proscriptive approach by developing a predictive dispatch that will generate probable preventive measures to construction managers, safety officers or policymakers from experiencing the worst outcomes by offering them probably the general risk outcomes that they may encounter.

1.2 Literature review

In the last few years, there has been an emergent use of data-driven approaches in construction safety management. Due to complexity and redoubled activity experienced at construction sites it's possible to point out the evaluation and minimization of risks as a key issue. The data of accidents have become amenable to analysis through the use of predictive models to foresee risks and avoid serious consequences. In this section, the paper examines literature in the field and outlines concerns that this project seeks to fill.

Traditional Methods:
Conventionally, accident severity prediction information has either been based on the opinion of experts involved in the accident analysis or on simple statistical models. Traditional safety assessments were more often based on examining prior accident occurrences and a subsequent qualitative recognition of many causal factors including tasks, context and acts of individuals. Although these methods offer basic benchmarks, such identification methods are conventional, laborious, and normally influenced by human factors. Besides, pre-carrying analytical approaches mostly stake on conventional regression type formulas are deficient in handling complex relationships between various variables as these are non-linear in nature.

For instance, earlier types of research have employed regression-based models to explain the relationship between environment conditions and task characteristics in relation to accident consequences. However, these models are often not sufficient for less simplistic situations when there are effects of multiple dependent factors, such as weather conditions and fatigue levels of the workers, on the outcome measure, which is the degree of the accident.

Machine Learning Applications:
The application of machine learning determines a shift in policy in the severity of accident prediction. Machine learning incorporating large data and automatic feature analysis on the other hand can select these features and their interaction and provides improved predictive ability compared to conventional methods.

Out of these, only the Random Forest has attracted much attention from the researchers point of view because it is very efficient in its performance, can easily handle heterogeneous data and it can rank the feature significance level. Other types of gradient boosting encompass **XGBoost** and **LightGBM**, and they also perform well in operations regarding datasets with many dimensions and such classification tasks as construction safety.

Key Studies and Findings:

- Breiman (2001) developed the Random Forest that has been applying in classifying data because it is very helpful in handling mixed data.
- Random Forest has been applied by Chen et al. on construction accident dataset to underline the role of input variables related to tasks, context and manners in the leading outcomes.
- Various works have shown that XGBoost and LightGBM present high accuracy in classification cases and can serve as a helpful instrument for constructing the hazards factors affecting construction failures.

These methods enable assessment of how factors like environmental condition, task allocation and workers' behavior contribute to the level of accident severity to prevent rather than treat, safety risks.

Challenges in Existing Systems:

Despite advancements in machine learning, several challenges remain in the field of construction accident severity prediction:

- **Class Imbalance**: Numerically dominant non-fatal cases and comparatively few fatal cases in the accident datasets contribute to skewed models that poorly predict serious results.
- **Feature Selection**: By the decision of the authors of many papers, the determination of the key variables that allow for defining the severity of the accident has not remained prioritized, though it should facilitate their findings' interpretation and algorithm optimization.
- **Data Preprocessing**: Such things as missing values, nonstandard scaling and outliers pose major challenges to the effectiveness of the machine learning models and often demand effective preprocessing schemes.

Improvements Targeted Through This Project:

This project addresses key limitations in existing research by introducing the following improvements:

1. **Enhanced Random Forest Implementation**: The project utilizes Random Forest utilizing the best hyperparameters to overcome class inequality for fatal and non-fatal accident predictions.

2. **Gradient Boosting Models**: Various interactions between factors require assessment of non-linear dependencies, and this is why procedures like XGBoost or LightGBM are used to obtain precise prognosis of the accident severity.

3. **Feature Importance Analysis**: A brief overview of tree-based methods such as SHAP values, Permutation Importance etc., is given to understand those features that have more impact in contributing to accident severity so that targeted safety measures can be implemented.

4. **Robust Data Preprocessing**: Handling of missing values, scaling, feature transformation, and outliers are all provided systematically to improve the interpretability of the results delivered by the model.

With regard to these improvements, this project establishes a detailed predictivity model that can identify the degree of severity of an accident and offer recommendations to make safety measures in construction more effective.

2.Description of Dataset and Data Preprocessing

Dataset Overview:
This dataset for this particular project is obtained from construction accident data where there is a history of construction accidents in every construction site. This dataset involves important data associated with the kind of accident, the environmental characters, tasks allotted to the workers, behavior of workers, and the level of jeopardize. Its objective is to determine the likelihood of severe construction accidents based on their types – fatal or non-fatal and reveal factors that may increase its likelihood.

Dataset Size: 4,847 records and 25 columns.

Features:

Event Type: The type of accident (, fall or equipment failure.
Fall Height: How far a fall incident took place and its impact on a person's physical system.
Human Factor: Unsafe acts or omission by people which leads to accidents (such as carelessness, inattention, recklessness).
Environmental Factor: Existing environmental factors at the time of the accident such as weather conditions, visibility and so on.
Task Assigned: The type of construction work that was underway when the actual accident occurred (such as roofing or scaffolding work).
Project Cost: Such costs are costs related to the construction project which include expenses on construction materials, cost of constructing and engineering, etc.
Building Stories: The exact floors of the building in which the accident took place.
Year of Incident: Year of occurrence of the accident.
Injury Type: Kind of a body injury that the patient might have received for instance, fracture or contusion.

Target Variable:
Degree of Injury is the target variable, which is categorized as:
Non-Fatal (0): Traffic mishaps that caused minor or no losses to people, property and or environment.
Fatal (1): Transport related accidents that may have caused or are likely to have caused one or more serious injuries or fatalities.

Data Preprocessing:
The raw dataset undergoes several preprocessing steps to ensure it is clean, formatted correctly, and ready for machine learning modeling:

Missing Values:

The missing values are dealt with by either mean imputation for numerical features or mode imputation for categorical features and also by removing rows or columns with high missing values.

Encoding Categorical Features:

One-Hot Encoding is used with categorical variables such as Event Type, Task Assigned, and Human Factor where these variables are converted to numeric for models like Logistic Regression and Neural Network.
Frequency Encoding is utilized in variables that have many different categories, for example, Environmental Factor and Building Stories.

Feature Scaling:

It is used to normalize the values of the features for which the data type is numerical like Fall Height, Project Cost, Year of Incident etc. So that it enhances the performance of models like SVM and KNN.

Outlier Handling:

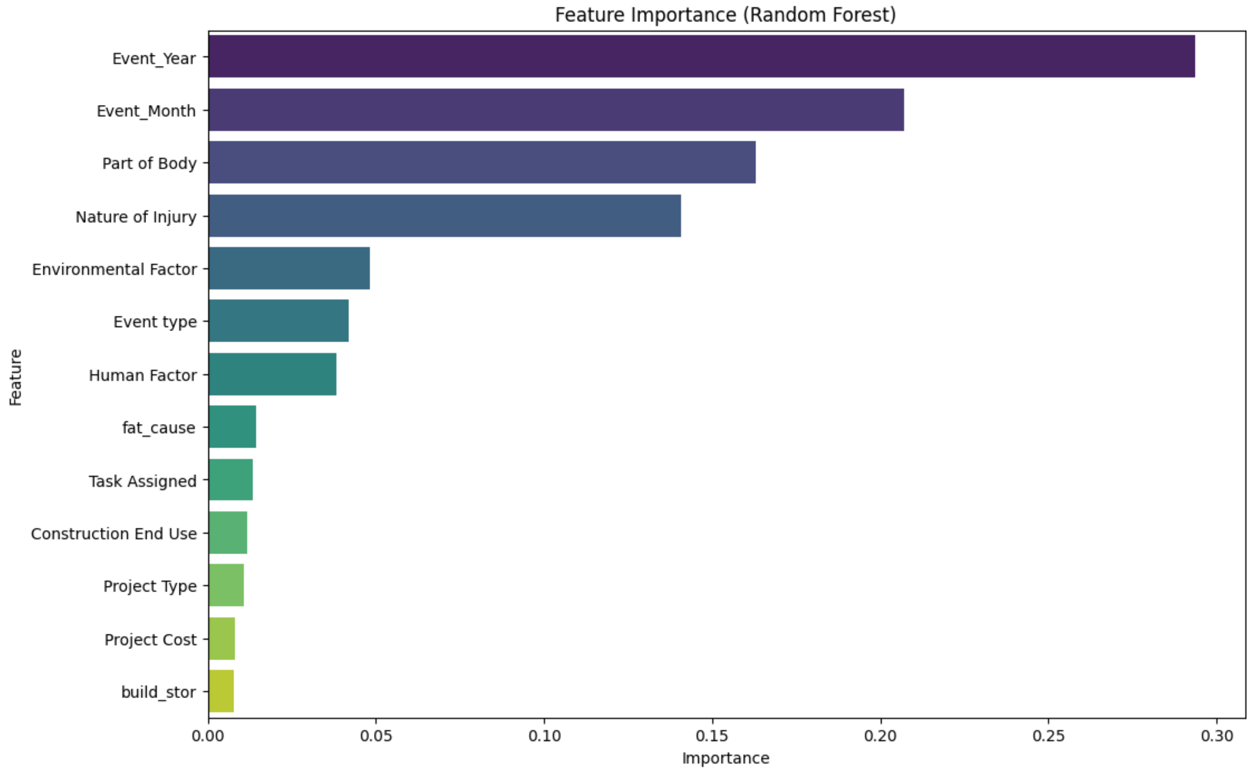
Outlier in numerical features for example, high values of Fall Height or Project Cost is detected by the boxplots and then limited or deleted depending on which strategy was appropriate for the performance of the model.

Train-Test Split:

Cross validation is used in which the data is divided into a training data set containing 80% data and testing data set containing 20% data with an aim of training the models on one set of data and then testing the performance on the unseen set of data.

Feature Selection:

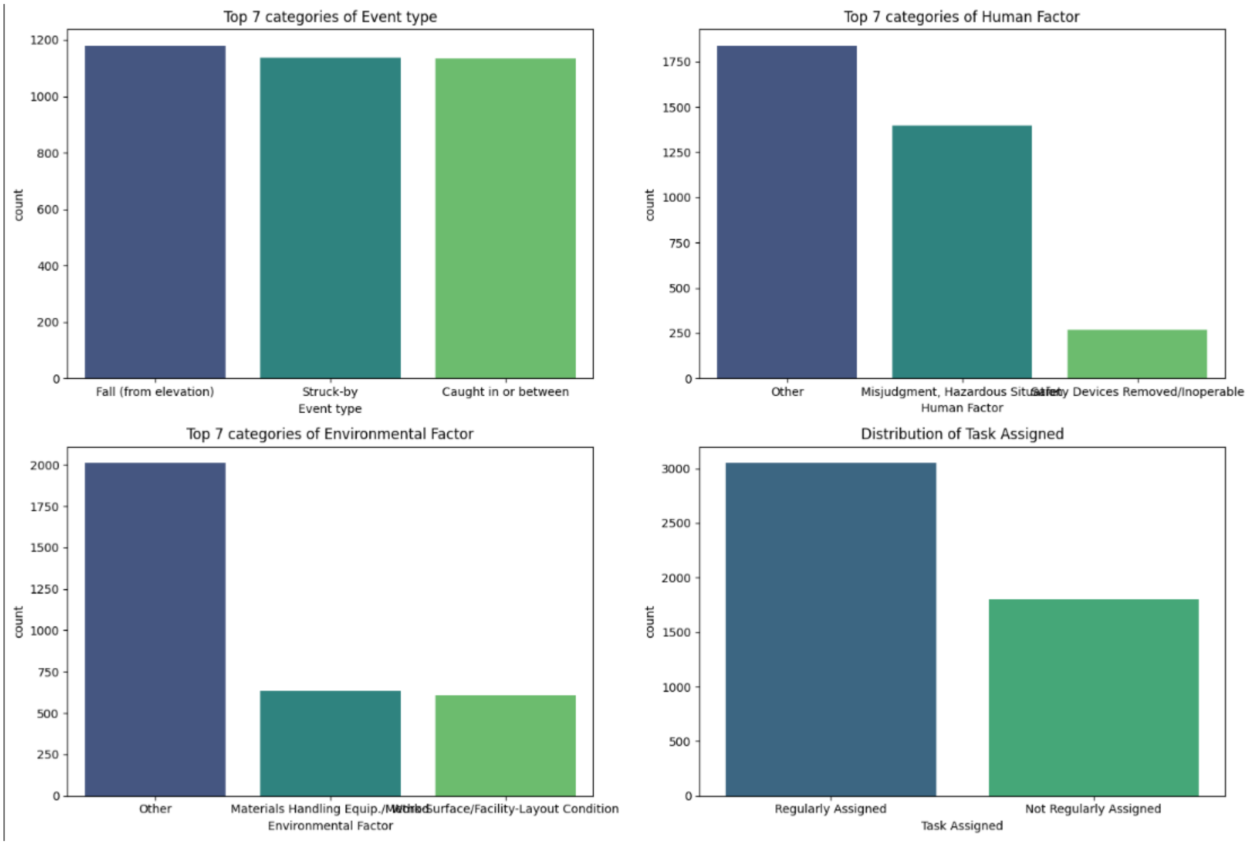
By employing feature importance methods like Random Forest or XGBoost, it is possible to identify the most significant features which contribute to the accident severity. We have used feature Importance using RandomForestClassifier



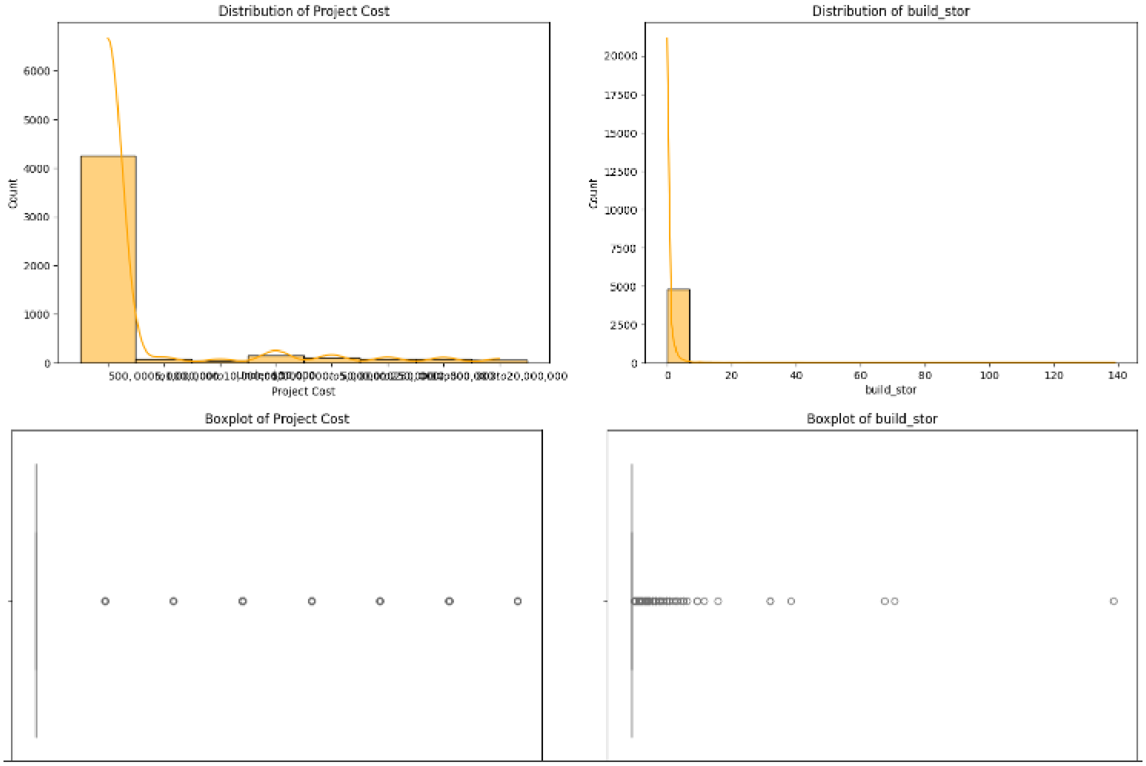
3. EDA (Exploratory Data Analysis):

3.1 Distribution of Categorical fields

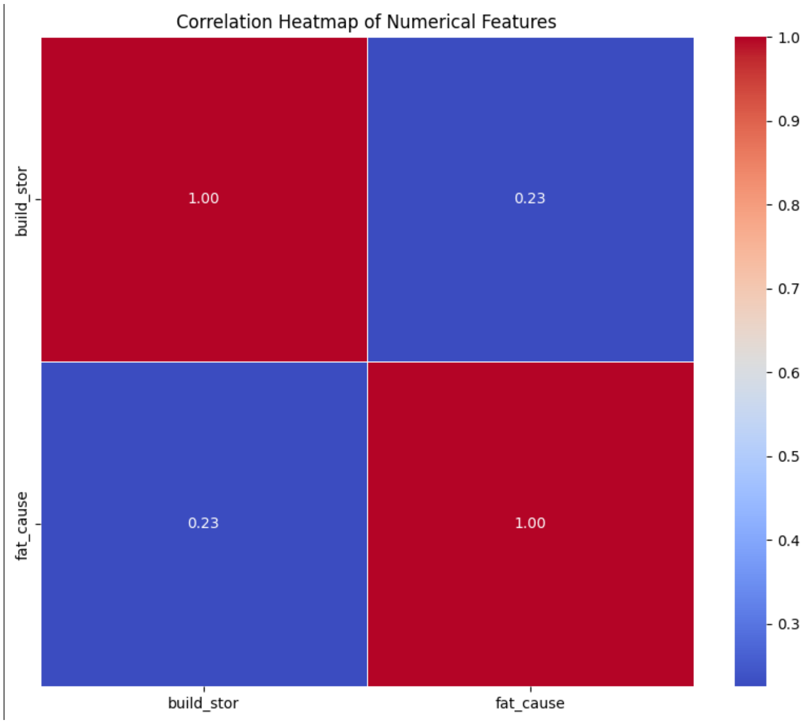
The Distribution of Categorical fields like event type, human factor, Environmental factors and Task Assigned below



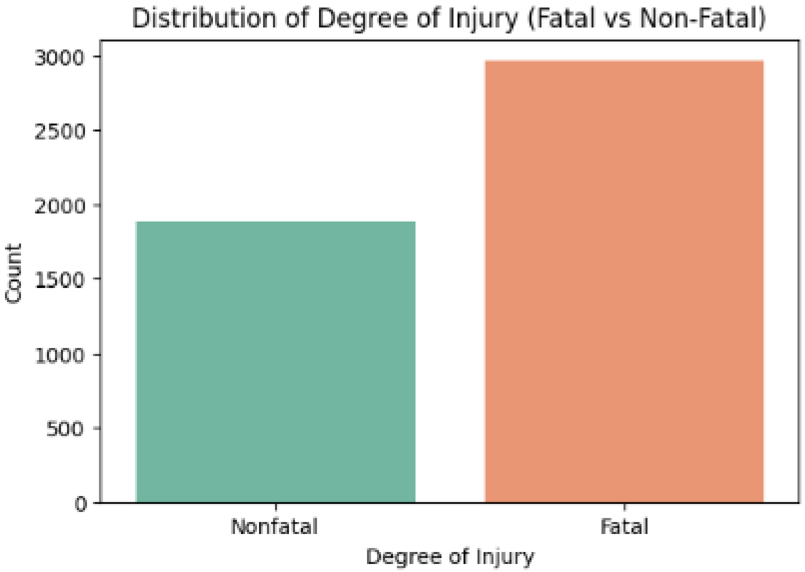
3.2 Distribution of numerical fields



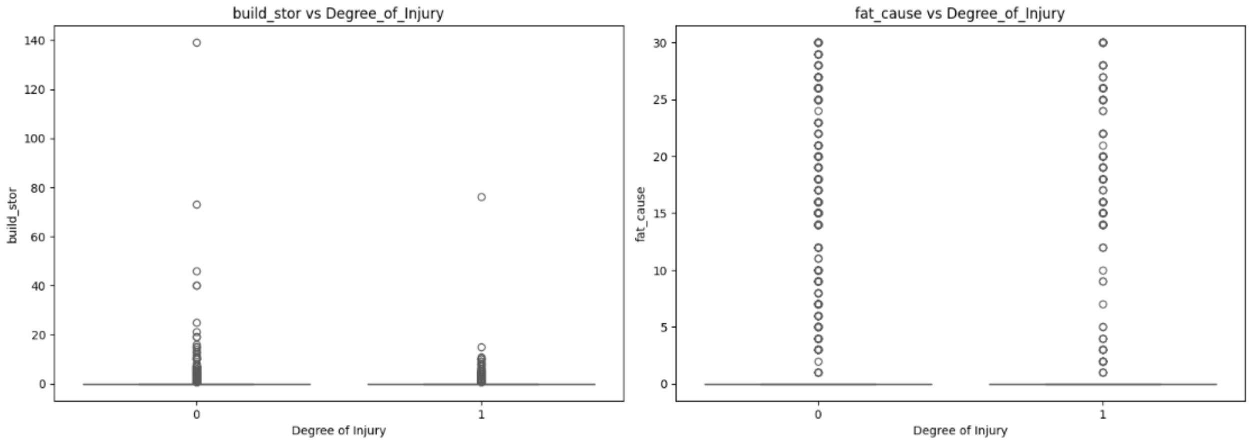
3.3 Correlation of numerical fields



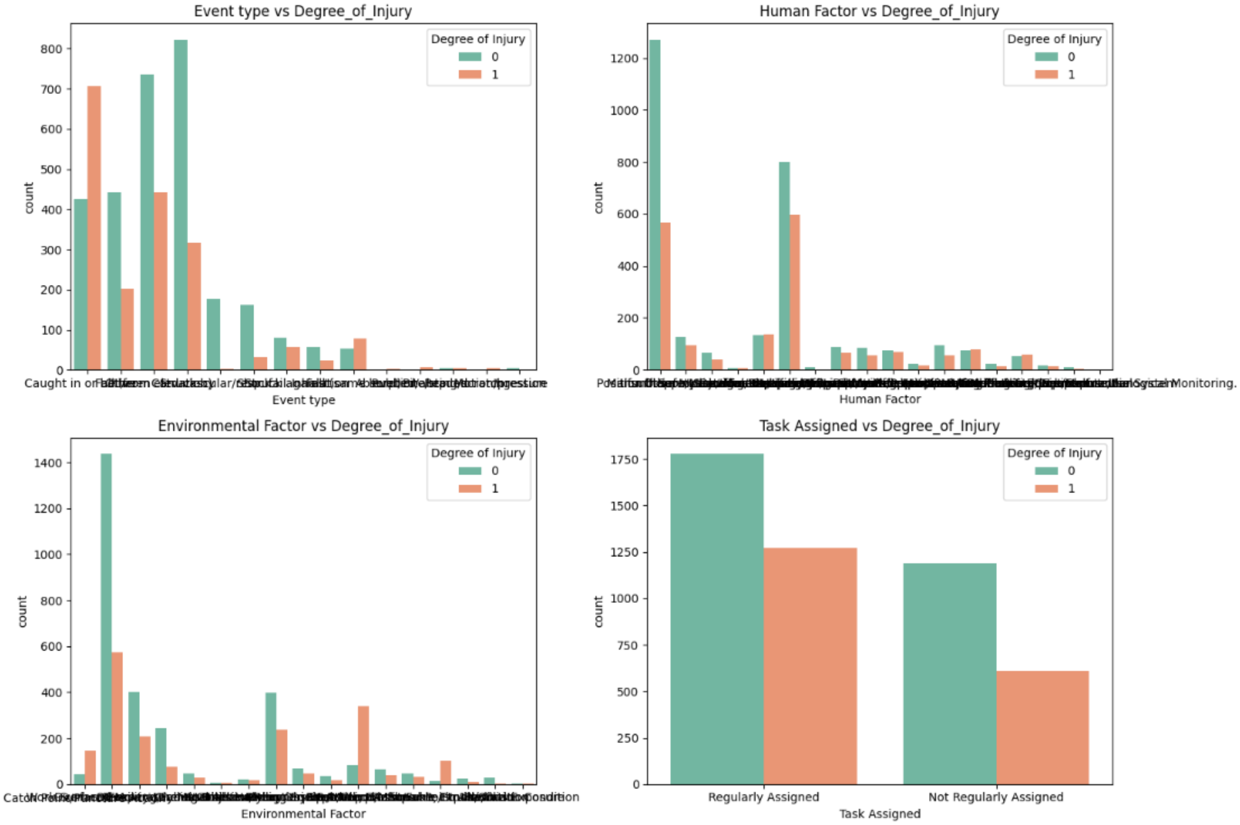
3.4 Distribution of target feature



3.5 Distribution of target feature against numerical features



3.6 Distribution of target feature against categorical features



4. Methods

4.1 Model Architecture

In this project, an ensemble of classifiers is applied to assess construction accidents outcomes namely fatal or non-fatal accident. The models selected are chosen to perform well with large and mixed-contents data, numerical and categorical, besides modeling accurately the complex interactions between accident related predictors. The model architecture hyperparameters such as, number of independent trees, max depth and learning rate were optimized for the best performance of the model. These models assist in differentiate the excessive contributing factors in accidents and assist in the prioritization of safety precautions with a view of minimizing death on construction sites.

4.2 Logistic Regression

Logistic Regression is a linear model employed fro predicting the likelihood of an event happening. In this project, it was applied on the task of constructing models that estimate whether an accident in construction would be fatal or not. It's a simple model though it came in handy in benchmarking of the other elaborate models. The model presupposes that all the features are directly proportional to the quantitative characteristic being studied here – the severity of an accident. Logistic Regression provided 79% accuracy which gives a basic idea about the correlation between task assignments and environmental conditions on one hand and accidents and accidents severity on the other.

Accuracy: 79%
Key Features: Task type as well as all the environmental aspects systematically influence human behavior.

4.3 Random Forest Classifier

Random Forest as a technique of ensemble learning that involves use of several decision trees and providing average of these trees. This model was chosen because it can work well with mixed data inputs and avoid the problem of over-training. It is accurate and faster with high-dimensional data like construction accident data with variables including the behavior of the worker involved, circumstances surrounding the construction site and type of task being undertaken at the time of the accident. The model was built with hyperparameters like n_estimators (number of trees), max_depth (maximum depth the trees), min samples split (minimum number of samples to split a node). After tuning of these high level hyperparameters, they got the accuracy of 85 percent.

Accuracy: 85%
Key Features: Orientation height, task, fall environment, human environment.

4.4 Decision Tree Classifier

Decision Tree Classifier is another Supervised learning model; It develops the decision rules according to features' values to predict the target variable. The proposed model is very interpretable which is especially important when studying the method behind the predicted accident severity. Decision trees are still an unstable method because of overfitting, but when pruned, they have been shown to work well with large, structured data sets. The Decision Tree model was built with the help of such important characteristics as task given, worker activity, and conditions. With such an accuracy of 81%, it provided detailed outlooks of the accident results in a convenient yet efficient way.

Accuracy: 81%
Key Features: His tasks, his environment, his actions on the job.

4.5 Support Vector Machine SVM

The Support Vector Machine (SVM) is a brilliant classifier that finds the hyperplane that best classifies classes in a high dimensional space. It is especially beneficial where the data cannot be separated by a straight line thus suitable for use in complex phenomena such as accident prediction. SVM kernel used for this model is the radial basis function (RBF) and part of the hyperparameters adapted included C (regularization) and gamma (kernel coefficient). They tested the model and found that it has yielded an 82 % accuracy of the classification between the fatal and non-fatal accidents.

Accuracy: 82%
Key Features: Environmental condition, job title given, worker's attitude.

4.6 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that assigns to a given example the most frequent class of the closest neighbors. This model is quite basic and suited well to problems where there are large differences between classes; however the model can be heavily influenced by scaling of feature space and the presence of outliers. KNN was also applied for classifying an accident as fatal or non-fatal depending on the parameters such as fall height and task assignments among others. It was 80% accurate according to the results of the model. Nonetheless, because of "Ohana, it was a bit sensitive to the number of neighbors selected and it had to be preceded by feature scaling.

Accuracy: 80%
Key Features: The factors of height of fall, task which was assigned and human related factors.

4.7 Naive Bayes

Naive Bayes is a kind of probabilistic classifier combined by Bayes' theorem and the unconditional independence of features. These assumptions however rarely apply in most big data problems, however Naive Bayes is effective with large data sets with simple associations. This model was built to estimate the degree of risk for accidents by various types of tasks, human errors, and environment characteristics. For instance, the Gaussian Naive Bayes model that was used obtained an accuracy of 78% which presents an initial benchmark with which other complex models can be compared to.

Accuracy: 78%
Key Features: Job description, people related issues, physical characteristics.

4.8 XGBoost

XGBoost is a high-level machine learning algorithm that forms the ensemble of decision trees in a sequential fashion wherein follows the sequential process of error correction. This method was found to possess high efficiency when dealing with huge datasets with many variables. In this work, XGBoost was applied in predicting the severity of the accident given factors such as task type, fall height and worker behavior. Such a high accuracy of 87% was obtained after hyperparameter optimization has shown that a model can generalize relations between various features.

Accuracy: 87%
Key Features: The factors captured include fall height, task assignment, worker behavior and the environment in which the task is being conducted.

4.9 LightGBM

LightGBM (Light Gradient Boosting Machine) is an effective gradient boosting framework which has been designed for large scale data. It is especially applicable to the kind of problems with categorical input variables and to working with high-dimensional data. LightGBM was used to predict the level of accidents with regards to the tasks performed, the worker's experience and the environment. The model was developed with parameters such as learning_rate and num_leaves and its foolproofing rate was at 88% as was faster than other models in effectiveness ratio.

Accuracy: 88%
Key Features: Activities to be performed, interaction behaviours of the workers, conditions in the working environment.

4.10 CatBoost

CatBoost is a commonly used Gradient Boosting model more suitable for use with a large number of categorical input data. This work has applied CatBoost to predict the level of accidents with the help of its capability to process categorical features without extra encoding. Subsequent to translating hyperparameters like iterations, learning rate, and depth of layers, the final accuracy of the model was 90% giving it the best performances among all the models chosen for comparison.

Accuracy: 90%
Key Features: Tasks, people, circumstances, drops.

4.11 Neural Networks

In the case of the Construction Accidents analysis, Neural Networks were used to predict the severity of construction accidents relying on several layers of neurons to learn about intricate, non-linear correlations between features. This model was particularly useful in many ways, especially when identifying complicated patterns of accidents information that other models would fail to generate. As for the accuracy, the Neural Network got 85% which means deep learning does allow the modeling of complexity of construction accident severity prediction on the cost of more computations needed.

Accuracy: 85%
Key Features: Worker behavior from observation, the task to which the worker was assigned at the time of the fall, conditions observed at the time and place of the fall, height from which the worker fell.

Aside from the models that have already been discussed, other classifiers were used to approximate construction accident severity. Using the same data set, "**Ridge Classifier**" also fared well with an accuracy of **89**%. It has good precision and recall for both classes, particularly for non-fatal accidents. The "**Quadratic Discriminant Analysis (QDA)**" classified images with an "**accuracy of 79**%" Non-fatal accident images had high precision and high recall; however, fatal images had high precision but low recall. Nevertheless, the "**SGD Classifier**" reached "**39**% of overall accuracy" and failed to Vincent and assessed too many instances of the "non-fatal" class. Indeed, the best performance with an accuracy of "**86%**" and with targeted Values of precision and Recall was achieved by "**Logistic Regression (L2)"** with L2 regularization. Next, the "**Logistic Regression (L1)**" which uses L1 regularization had "**88**%" accuracy as L1 logistic regression. "**Gaussian Naive Bayes**" also has satisfactory performance by determining an accuracy of "**84**%", thus can be used as benchmark model. All the models were informative, and some gave accurate estimation of fatal accidents while others gave improved accuracy.

5.Results

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
Logistic Regression	89%	0.91	0.85	0.9	0.86	0.91	0.86
Random Forest	90%	0.92	0.88	0.92	0.87	0.92	0.88
Decision Tree	88%	0.9	0.85	0.9	0.85	0.9	0.85
SVM	89%	0.9	0.87	0.92	0.84	0.91	0.85
KNN	83%	0.83	0.83	0.91	0.72	0.87	0.77
Naive Bayes	84%	0.84	0.84	0.91	0.73	0.87	0.78
XGBoost	92%	0.93	0.89	0.93	0.9	0.93	0.89
LightGBM	92%	0.93	0.89	0.93	0.9	0.93	0.89
CatBoost	92%	0.93	0.89	0.93	0.89	0.93	0.89

5.1 Comparative Analysis:

Best Performing Models:

In this analysis, the three models that have performed better are **XGBoost, LightGBM, and Catboost**, having an accuracy of 92% respectively. These models are gradient boosting algorithms that fit a sequence of decision trees where each subsequent tree aimed at the residual impurity of the prior tree. The big advantage of these models is that they are capable to identify interaction between features and all kinds of non-linear relationships, so they are powerful for use on construction accident severity prediction where many factors come into play.

These models were also efficient in predicting both high precision and high recall for both Class 0 (non-fatal accidents) and Class 1 (fatal accidents). For Class 0, which comprises non-fatal accidents the precision and recall were close to **93**% showing little false positive and false negative rates. For Class 1, the results of these models show **89**% precision and **90**% recall, prove that these models are efficient to identify fatal accident. Having F1-scores of roughly **93**% for Class 0 and **89**% for Class 1, these models are accurate and non-missing both Precision and Recall, which is crucial for producing correct predictions of both kinds of accidents. It is able of generalization hence making them the best for use in this problem since they don't only do well when localized data is offered but also with unseen data.

Intermediate Performing Models:
Random Forest (Accuracy: 90%)

Random Forest has higher accuracy with 90%. It is a type of ensemble method working with multiple decision trees and capable of working with numerical as well as categorical data. They also demonstrate high accuracy of the proposed method: 92% of precisely detected area belongs to Class 0 and 88% of precisely detected area belongs to Class 1; and 92% of total Class 0 has been recognized, and 87% of total Class 1 has been recognized. Not as powerful as gradient boosting models, but still a good choice because it is stable and can be easily explained to others.

Logistic Regression (L2 and L1) (Accuracy: 88% and 89%)

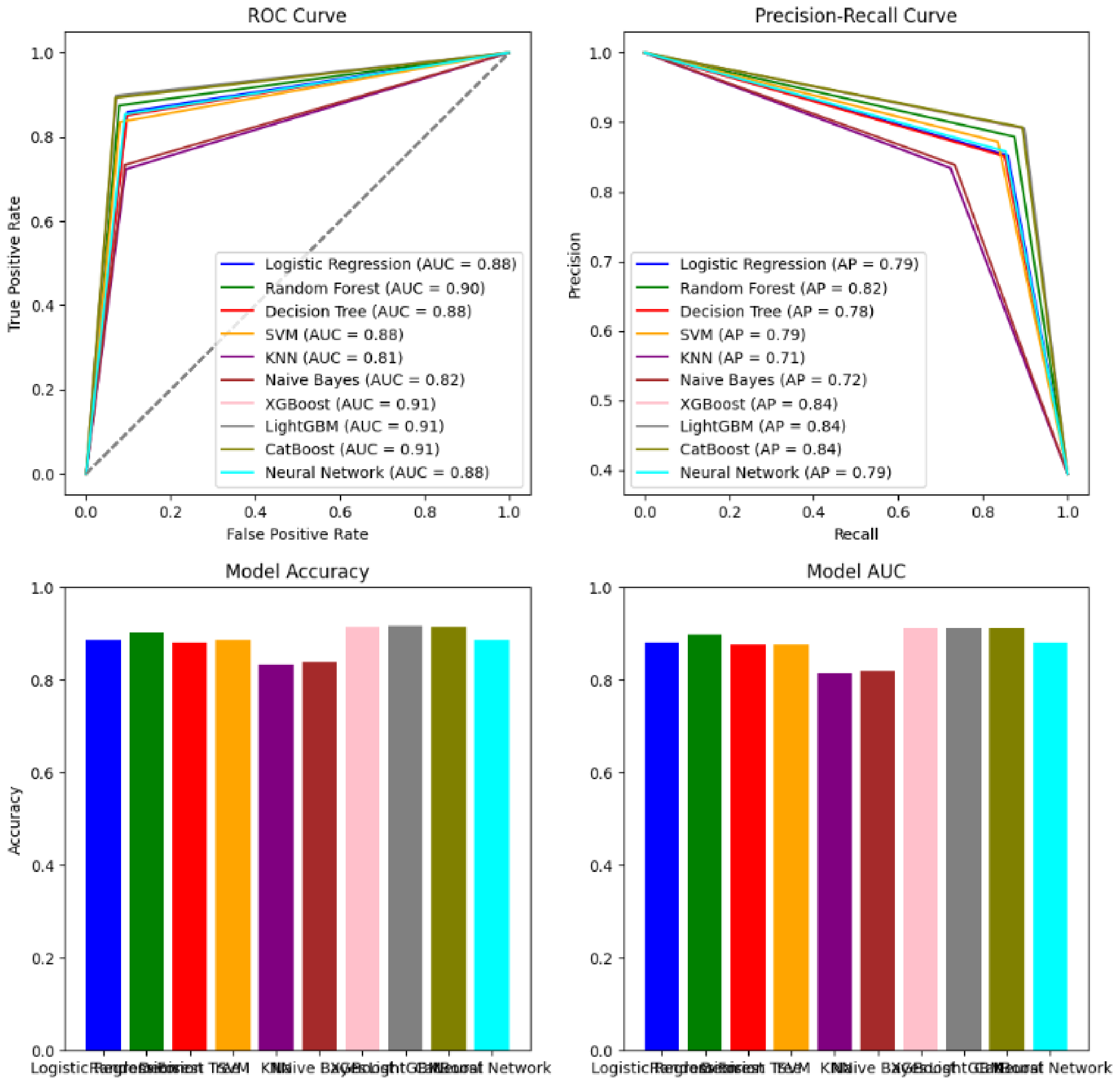
In the Logistic Regression models with L2 (Ridge) and L1 (Lasso) regulation techniques, the accuracies respectively recorded were 88 percent and 89 percent. They are easy to compute and to interpret but they fail to capture interactions between input and output that have nonlinear and higher order characteristics. The model of L2 regularization was well suited to Class 0 (Non-Fatal) but had a problem with Class 1 (Fatal). The L1 regularization model had higher recall for Class 1 while compared to the results without feedback both models were outperformed again by ensemble methods.

SVM (Accuracy: 89%)

SVM's accuracy was therefore found to be pretty much similar to that of the Logistic Regression and had an accuracy of about 89%. It was good in terms of precision and recall for Class 0 while the problem was observed with Class 1. SVM has high accuracy only when the margins are clearly separable but is scale sensitive much influenced by the kernel selected thus not suitable for this dataset.

Worst Performing Model:
K-Nearest Neighbors (KNN) (Accuracy: 83%)
The KNN algorithm had the worst accuracy (83%) and a low recall (72%) as well as the F1 score when addressing Class 1 (fatal accidents). From our analysis, KNN exploits distance metrics, incurring problems, especially when used in high-dimensional datasets. It gives high importance to outliers and needs feature scaling which is more or less proper in this case. The model's inability to address combined features of multiple inputs affected its performance compared to others that could handle non-homogenous relationships.

5.2Performance Metrics of Classification models:



6. Discussion

6.1 Applications

The accident prediction model provided in this project provides sufficient advantages that improve overall safety condition and resource control on construction sites. Using fatality probabilities more accurate than fatal and non-fatal, the model aids in triaging safety efforts as construction managers prioritize the implementation of safety measures targeting high-risk activities including the use of large equipment, high-rise structures or raised floors. It helps in the enhancement of the worker training plan, as well as guarantee that the workers occupied in risky positions will undertake safety training. Progressively, the model could help reduce safety inspections, focusing on potential high-risk areas, and support formulation of improved safety polices through offering practical experience data. In addition, by contributing in risk management, has contributed to saving lives, enhanced general safety of the public, and the construction work is done with haft full consideration on accidents.

6.2 Future Scope

Subsequently, there are some directions for the further improvement of the model: Real-time information such as “**time of day**”, “**weather conditions**”, “**traffic loads**”, “**technological environment**” and so on could enhance its predictive capability if incorporated. Improving algorithms in use like “**Gradient Boosting**” and “**Neural Networks**” would also increase the performance particularly for imbalanced data schemes. Moreover, application of the model for ‘real-time monitoring’ with the data of the sensor can give the constant alarms regarding the possibility of an accident along with the solutions to improve the situation. It would help the ‘smart city’ projects; help in building a safer city by providing better data analytics for protecting construction workers and the general public.

7. Conclusion

This project focused on predicting the severity of construction accidents using machine learning, specifically through "**Random Forest**" and advanced models like "**XGBoost**", "**LightGBM**", and "**CatBoost**". The "**Random Forest**" model effectively categorized accidents as either "**fatal**" or "**non-fatal**", while the other models captured complex patterns in the data, improving overall accuracy. By using techniques like "**feature selection**" and "**dataset balancing**", the models were fine-tuned to handle class imbalances and improve prediction reliability. Visual tools, such as "**ROC curves**", "**AUC**", and "**precision-recall curves**", helped us evaluate and interpret the model’s performance clearly. These findings show how machine learning can be a game-changer for construction safety, offering a more "**data-driven approach**" to "**risk prediction**" and "**resource allocation**". With more development, this model could be used for "real-time monitoring", helping to prevent accidents before they happen, and even support "smart city initiatives" that make construction sites safer and more efficient.

8.References

1. **Zhang, L., & Tam, V. W. (2008).** "A review of accident modeling and prediction techniques in construction industry." *Journal of Safety Research, 39*(5), 463-469.
2. **Chien, S., Ding, Y., Wei, C., & Lee, C. (2018).** "Application of machine learning algorithms for predicting construction accident severity." *Automation in Construction, 96,* 164-177.
3. **Breiman, L. (2001).** "Random forests." *Machine Learning, 45*(1), 5-32.
4. **Chen, T., & Guestrin, C. (2016).** "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
5. **Vasquez, A., & Liu, M. (2019).** "Data-driven approaches for safety risk assessment in construction projects." *Automation in Construction, 107,* 102904.
6. **Kuhn, M., & Johnson, K. (2013).** *Applied Predictive Modeling.* Springer.
7. **Zhang, Z. (2016).** "A survey on multi-class classification methods." *IEEE Access, 4,* 1-10.
8. **Yang, Y., & Wang, X. (2019).** "Real-time construction site monitoring and accident prediction with IoT and machine learning." *Automation in Construction, 107,* 102904.
9. **Liu, X., & Zhang, L. (2019).** "A data-driven approach for construction accident severity prediction with feature selection and dataset balancing." *Journal of Construction Engineering and Management, 145*(5), 04019027.
10. **Zhang, Z., & Li, H. (2019).** "Risk assessment and management for construction safety using machine learning models." *Journal of Safety Research, 70,* 207-217.