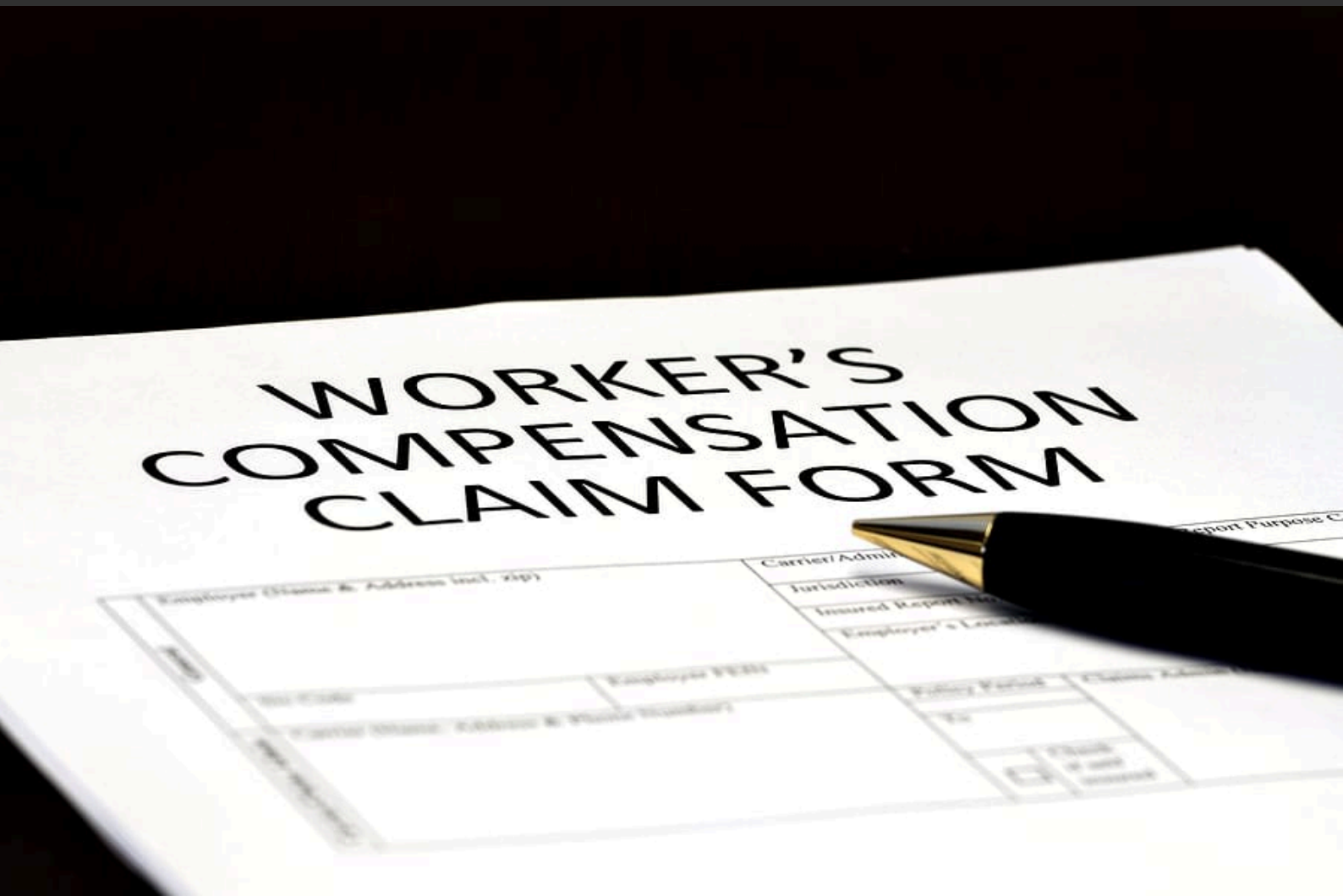
SEPTEMBER, 2024

TO GRANT OR NOT TO GRANT: DECIDING ON COMPENSATION BENEFITS

GROUP PROJECT
MACHINE LEARNING 2024/2025
DUE: DECEMBER 23RD (11:59 AM)



01

I. INTRODUCTION

The New York Workers' Compensation Board (WCB) administers and regulates workers' compensation, disability, volunteer firefighters, volunteer ambulance workers, and volunteer civil defence workers' benefits. As the regulating authority, the WCB is responsible for assembling and deciding on claims whenever it becomes aware of a workplace injury. Since 2000, the WCB has assembled and reviewed more than 5 million claims.
However, manually reviewing all claims is an arduous and time-consuming process. For that reason, the WCB has reached out to Nova IMS to assist them in the creation of a model that can automate the decision-making whenever a new claim is received.

II. PROJECT GOALS

The goal of your project is three-fold:

1. **Multiclass Classification Benchmarking:** Create a classification model that can accurately predict the WCB's final decision on what type of injury (***Claim Injury Type***) should be given to a claim. To do that, the WCB has provided labelled data with all claims assembled between 2020 and 2022. You will need to **develop a consistent model assessment strategy** that will allow you to **create and compare different candidate models to find the most generalizable one**.
2. **Model Optimization:** After selecting your best (or set of best) models, you are encouraged to explore ways to improve their performance (e.g. hyper-parameter tuning or pre-processing/feature selection adjustments). **You should compare the optimized model with your previous models and discuss your findings**.
3. **Additional Insights: This project segment is open-ended**, meaning you can explore as many ideas as you desire (as long as you make them explicit and understandable). Here are **some possible suggestions**:
    a. ***Analyze and discuss the importance of the features for the different values of the target variable.***
    b. ***Create an analytics interface that returns a prediction when new inputs are given.***
    c. ***Create a model that predicts other variables ('WCB Decision' or 'Agreement Reached') and check whether using these variables as features improves the performance of your models.***

02

III. DATASET

You have access to two different datasets:

In the **training set**, you will find the claims data assembled **from the start of 2020 till the end of 2022**. You will have features and three specific ground truths associated with each assembled claim. Use the training data and its features to build and validate your machine-learning models.

In the **test dataset**, you can still access the same descriptive attributes associated with the claims assembled from January 2023 onward. However, you will not have access to information that would only exist after a decision about these claims has been made. You must use the models trained with the training set to predict ***Claim Injury Type*.** You will be able to know how well your model performs on this data through a Kaggle competition.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
Accident Date	Injury date of the claim.
Age at Injury	Age of injured worker when the injury occurred.
Alternative Dispute Resolution	Adjudication processes external to the Board.
Assembly Date	The date the claim was first assembled.
Attorney/ Representative	Is the claim being represented by an Attorney?
Average Weekly Wage	The wage used to calculate workers' compensation, disability, or an Paid Leave wage replacement benefits.
Birth Year	The reported year of birth of the injured worker.
C-2 Date	Date of receipt of the Employer's Report of Work-Related Injury/Illness or equivalent (formerly Form C-2).
C-3 Date	Date Form C-3 (Employee Claim Form) was received.
Carrier Name	Name of primary insurance provider responsible for providing workers' compensation coverage to the injured worker's employer.
Carrier Type	Type of primary insurance provider responsible for providing workers' compensation coverage.
Claim Identifier	Unique identifier for each claim, assigned by WCB.

03

ATTRIBUTE	DESCRIPTION
County of Injury	Name of the New York County where the injury occurred.
COVID-19 Indicator	Indication that the claim may be associated with COVID-19.
District Name	Name of the WCB district office that oversees claims for that region or area of the state.
First Hearing Date	Date the first hearing was held on a claim at a WCB hearing location. A blank date means the claim has not yet had a hearing held.
Gender	The reported gender of the injured worker.
IME-4 Count	Number of IME-4 forms received per claim. The IME-4 form is the "Independent Examiner's Report of Independent Medical Examination" form.
Industry Code	NAICS code and descriptions are available at: https://www.naics.com/search-naics-codes-by-industry/.
Industry Code Description	2-digit NAICS industry code description used to classify businesses according to their economic activity.
Medical Fee Region	Approximate region where the injured worker would receive medical service.
OIICS Nature of Injury Description	The OIICS nature of injury codes & descriptions are available at https://www.bls.gov/iif/oiics_manual_2007.pdf.
WCIO Cause of Injury Code	The WCIO cause of injury codes & descriptions are at https://www.wcio.org/Active%20PNC/WCIO_Cause_Table.pdf
WCIO Cause of Injury Description	See description of field above.
WCIO Nature of Injury Code	The WCIO nature of injury are available at https://www.wcio.org/Active%20PNC/WCIO_Nature_Table.pdf
WCIO Nature of Injury Description	See description of field above.
WCIO Part Of Body Code	The WCIO part of body codes & descriptions are available at https://www.wcio.org/Active%20PNC/WCIO_Part_Table.pdf
WCIO Part Of Body Description	See description of field above.
Zip Code	The reported ZIP code of the injured worker's home address.
Agreement Reached	**Binary variable:** Yes if there is an agreement without the involvement of the WCB -> **unknown** at the start of a claim.
WCB Decision	**Multiclass variable:** Decision of the WCB relative to the claim: "*Accident*" means that claim refers to workplace accident, "*Occupational Disease*" means illness from the workplace. -> **requires WCB deliberation** so it is unknown at start of claim.
Claim Injury Type	**Main target variable:** Deliberation of the WCB relative to benefits awarded to the claim. **Numbering indicates severity**.

04

IV. OUTLINE

Your project deliverables (especially the report) should respect the following outline:

Group Member Contribution

Abstract
A small summary of your work (200 to 300 words). The abstract should give an overview of your work: What is the context? What are your goals? What did you do? What were your main results, and what conclusions did you draw from them?

I. Introduction
- Overview of the project
- Main goals of the project
- Are there any similar works/applications? What has been done? What did other researchers find? What would you expect your results to be based on their previous findings?

II. Data Exploration and Preprocessing
- Description of data received -> key insights
- Steps taken to clean and prepare the data

III. Multiclass Classification
- Additional preprocessing steps adopted
- Feature Selection Strategy
- Explanation of model assessment strategy and metrics used
- Comparison of performance between candidate algorithms
- **Optimization efforts**: presentation, results and discussion

IV. Open-Ended Section
- Objectives for the Section
- Description of the actions taken
- Results and discussion of main findings

V. Conclusion
- Summary of initial objectives and discussion of corresponding findings
- Do the findings match what you initially expected? How?
- Discussion of limitations of your work (e.g. what could you have done differently)
- Suggestions for possible work to follow on your work.

V. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A Jupyter notebook (or a zipfile) featuring all the code you used throughout the project to:
  a. Decide on your final solution
  b. Obtain your final results (code that helped you make decisions but does directly contribute to reach should be included, but commented).
- A report that describes the analytical processes and the conclusions obtained with, at most, 10 pages (excluding cover, abstract and annexes).
  - The file naming format should follow ***Machine_Learning_GroupXX_Report.pdf***, where ***GroupXX*** should be your group number. The report should follow these settings**:**
    - **Heading 1: Calibri, Size 14 pt, in bold**
    - **Heading 2 (if needed): Calibri, Size 13 pt, in bold**
    - **Text: Calibri, Size 11 pt, line spacing of 1.15 pt and paragraph spacing of 6 pt**
  - The body of text should only include Figures and Tables that are essential to understanding your work. Supporting figures and Tables can be added to Annexes.
  - Please make sure all figures and Tables (including the ones in annexes) are identified and referenced in the text. Any figure or table should have an explicit purpose to be included.

VI. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Report Quality and Storytelling	10	2
Data Exploration & Initial Preprocessing	15	3
Multiclass Classification	40	8
Open-Ended Section	25	5
Conclusion	10	2
Extra Point: Have Project Be Publicly Available on GitHub	-	1

06

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken into account for each topic:

- **Report Quality and Storytelling (2v):** A good report should, by itself, give the reader a clear picture of the problem you are tasked with, the steps you took, the rationale behind those steps, your main results and your insights. When referencing a figure, ensure you direct the reader's attention to the point you want to convey. This section also encompasses the overall quality of your introduction and conclusions.
- **Data Exploration & Initial Preprocessing (3v):** Describe the data and extract meaningful insights that you consider helpful. Avoid adding visualizations and elements that do not address the problem at hand. In addition, it should also unambiguously explain the steps and rationale behind your steps into cleaning the data into something usable by your predictive models.
- **Multiclass Classification (8v):** Describe your strategy for the text classification objective. This section is separated into different components:
  - Kaggle Performance: 2v
  - Additional Preprocessing: 1v
  - Feature Selection: 1v
  - Modelling approach - model assessment strategy (holdout, cross-validation, etc...) and algorithms (minimum of 5 covered in class) used: 1.5v
  - Performance assessment - rationale for choice of evaluation metric(s) and interpretation of results: 1.5v
  - Model optimization: 1v
- **Open-Ended Section (5v):** Describe your strategy for the additional insights objective. This section is separated into different components:
  - Formulation and Adequacy of the Objectives: 0.5v
  - Difficulty of tasks: 1.5v
  - Correctness/efficiency of implementation: 1.5v
  - Discussion of results: 1v
  - Alignment between results and communicated objectives: 0.5v
- **Conclusion (2v):** A good conclusion perfectly summarizes the work done. It draws from the information and questions formulated the introduction and directs the results obtained to address them. Moreover, it also lays out the path ahead, such as discussing the limitations of the work and hinting at what could be done in the future.

07 VII. PARTING NOTES

1. Deliveries **after the deadline** will be **penalized at 1 point per day.**
2. Deliveries made **before the deadline** will **receive a bonus of 0.15 points per day of delivery in advance (**up to a **maximum of 1 point).**
3. For modelling purposes, **using Lazy Predict or similar AutoML (e.g. Feature Tools, TSFresh) packages is explicitly off-limits and will result in a 1-point penalty.**
4. The report will be the primary method of evaluating your work. When preparing it, remember that a reader should be able to understand your work without needing to check your notebook. We won't consider any steps or results not mentioned in your report.
5. Everything in your report must have a clear purpose. Avoid including irrelevant, unimportant, or redundant information, as the space is limited, and you will need it. Also, please **don't provide theoretical explanations of topics covered in class**.
6. The trustworthiness of the information you provide is key. You should look to **source information from peer-reviewed journals** (thus, avoid citing Medium, TowardsDataScience and similar sources).
7. Before submitting, **run your notebook from the start one last time** (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
8. Your submitted notebook should include all the unneeded code you used to obtain your final solution**, but it should also be commented.**
9. We **will run your Jupyter Notebooks if we have any doubts**. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
10. The report and code will pass through a process of plagiarism and AI generation checking.
11. **You must submit to the Kaggle competition to get points for that component**.
12. When determining the grade for your work, there will be a comparative component between it and the work presented by your peers.

Friendly Reminders:

1. Attendance at the defense is mandatory for approval in the project. The defense has a group component and an individual component.
2. As questions are individualized, every group member should be able to understand what was done at every step of the way.
3. **If something is good enough to be mentioned in the report, it is also good enough to know. DO NOT include techniques/algorithms/steps you cannot explain in your report: we may (and probably will) ask about them in the defense.**
4. **Finished is better than perfect.**