Technical Exercise | Suncorp

**Workers' Compensation Claims Estimates**

A Data-Driven Approach to Process Improvement

horizontal line

# 

# Introduction

In the realm of Workers' Compensation, efficient and accurate claims processing is crucial for maintaining financial stability and ensuring customer satisfaction. Suncorp recognizes the importance of leveraging Data Science to enhance the claims management processes. As a Data Scientist, my primary objective is to identify and implement opportunities to improve the accuracy and efficiency of their claims handling procedures.

The current manual process employed by their claims handlers involves extensive time and effort to review claim characteristics and estimate costs. These estimates significantly impact their financial reserves and must be precise and consistent. Unfortunately, the existing system often falls short in these areas, necessitating frequent adjustments. Additionally, the identification and triaging of complex claims pose significant challenges, often resulting in delays and suboptimal handling.

To address these issues, I have undertaken a comprehensive analysis of historical claims data, aiming to develop a data-driven solution that enhances their claims estimation accuracy and streamlines the identification of complex claims. This initiative is supported by a dataset provided by the business, which contains valuable insights from past claims.

In this report, I will outline the steps taken to analyze the dataset, develop predictive models, and derive actionable insights. The objective is to showcase how Data Science methods can be applied to revolutionize their claims processes, thereby improving efficiency, accuracy, and overall service quality. The findings and recommendations will be presented to the business owner and their technical advisor, highlighting the potential impact and benefits of the proposed solutions.

By leveraging advanced analytical techniques and machine learning models, I aim to transform their claims handling approach, ensuring timely and accurate estimations while efficiently managing complex claims. This endeavor underscores their commitment to innovation and continuous improvement in delivering superior Workers' Compensation services at Suncorp.

# Approach

## Task 1

*Our claims handlers spend a lot of time going through the claims characteristics and put cost estimates in the system. It is important that we get accurate estimates because these impact our reserves. Our estimates today are not very accurate and consistent and we have to modify them regularly.*

To address the challenges highlighted by the business owner regarding the accuracy and consistency of claim cost estimates in the Workers' Compensation business, we can approach this by developing a predictive model using the historical claims data provided.

### Here’s a plan to tackle the problem:

#### Step 1: Data Exploration and Preprocessing

* **Loading Data**: We'll begin by loading the dataset to understand its structure and contents.
* **Exploratory Data Analysis (EDA)**: This involves summarizing main characteristics of the data, looking for missing values, outliers, and understanding the distribution of various features.
* **Feature Engineering**: Based on the insights from EDA, we can engineer new features that might help improve the model's performance.
* **Data Cleaning**: Handling missing values and anomalies based on the EDA findings.

#### Step 2: Visualization

* **Data Distributions**: Visualizing the distribution of various features like Age, Weekly Wages, Hours Worked Per Week, etc.
* **Correlation Analysis**: Checking how the different features correlate with the Ultimate Incurred Claim Cost.
* **Trends and Patterns**: Identifying any trends that might influence claim costs.

#### Step 3: Predictive Modeling

* **Model Selection**: Considering models suited for regression tasks such as Linear Regression, Random Forest, Gradient Boosting Machines (GBM), and Neural Networks.
* **Training**: Splitting the data into training and validation sets to train the models.
* **Model Evaluation**: Using metrics like RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) to evaluate model performance.

#### Step 4: Business Insights and Recommendations

* **Model Interpretation**: Using tools like SHAP or LIME to interpret the model results in a business-friendly manner.
* **Implementation Strategy**: Discuss how the model can be integrated into the existing claims processing workflow.
* **Continuous Improvement**: Establishing a feedback loop to continually refine the model as more data becomes available.

## Task 2

*It is very hard to identify complex claims and triage them to the right people. It would be nice if we can detect these complex claims early and get them to the right claims handlers to manage.*

The objective of this analysis is to improve the existing claims processes within the Workers' Compensation business at Suncorp by leveraging data science methods. One of the primary challenges identified is the early detection of complex claims to ensure they are triaged to the right claims handlers. This report outlines the approach and findings in building a predictive model to identify such complex claims based on historical claims data.

Here's a step-by-step plan:

#### Step 1: Data Exploration and Preprocessing

Data Preprocessing:

* Converted date columns to datetime format and created additional features from them.
* Handled missing values and encoded categorical variables.

Exploratory Data Analysis (EDA):

* Conducted summary statistics and visualized the distribution of key features.
* Generated a correlation heatmap to understand relationships between features and the target variable.

#### Step 2: Visualization

* Visualize the distribution of key features.
* Analyze relationships between features and the target variable (UltimateIncurredClaimCost).

#### Step 3: Predictive Modeling

* Split the data into training and testing sets.
* Train regression models to predict UltimateIncurredClaimCost.
* Evaluate model performance using appropriate metrics.

#### Step 4: Business Insights and Recommendations

* Identifying important features driving the predictions.
* Visualize model performance and feature importance.
* Create visualizations to aid business understanding and decision-making.

#### Step 5: Complex Claim Detection:

* Defined a threshold for complexity based on the 90th percentile of UltimateIncurredClaimCost.
* Flagged claims exceeding this threshold as complex.
* Evaluated the model's performance using classification metrics.

### Results

Data Preprocessing:

* Successfully handled missing values and created new features for better prediction.
* Encoded categorical variables to numerical format.

Exploratory Data Analysis:

* Identified that InitialIncurredClaimCost has a strong correlation with UltimateIncurredClaimCost, indicating it as a significant predictor.

Model Performance:

* The Random Forest Regressor was chosen for its superior handling of non-linearities and interactions.
* The complexity threshold was set at the 90th percentile of the UltimateIncurredClaimCost.

Complex Claim Detection:

* Claims exceeding the defined threshold were flagged as complex.
* The model's ability to detect complex claims was evaluated using a confusion matrix and classification report, yielding good accuracy metrics.

The predictive model developed in this analysis provides a robust mechanism to identify complex claims early in the process. By flagging high-cost claims, the business can triage them to specialized claims handlers, improving overall efficiency and outcomes. Future work can focus on refining the model further and integrating additional data sources for enhanced prediction accuracy.