PROJECT PROPOSAL: PREDICTING INSURANCE CLAIM SEVERITY USING MACHINE LEARNING

Prepared by:
Rebecca Stalley-Moores

Project Proposal: Predicting Insurance Claim Severity Using Machine Learning

---

Prepared by: **Rebecca Stalley-Moores**

Date: **28/05/2025**

---

1. Background and Motivation

In the insurance industry methods of improving risk assessment and claims management are advancing at an unprecedented rate. At the forefront of such innovations are companies like Accelerant which leverage advanced analytics and machine learning to drive intelligent policy strategies.

The ability to predict the severity of insurance claims is essential for accurately pricing policies and managing reserves, which, in turn, enables insurers and MGAs to maximise profitability and design bespoke coverage for speciality insurance markets.

This project aims to build and compare multiple models to predict claim costs, illustrating how data-driven insights can enhance risk assessment and pricing accuracy.

---

2. Objectives

The primary objective of this project is to develop and compare multiple models capable of predicting the severity of insurance claims based on historical claim data, selecting the most accurate and interpretable model for final deployment. By quantifying expected claim costs, the model will support more accurate policy pricing and improved reserve planning.

Secondary goals include:

- Identifying the most influential factors affecting claim severity
- Evaluating and comparing multiple regression-based and ensemble learning techniques

3. Methodology

3.1. Data Collection

This project will use the publicly available Allstate Claims Severity dataset, which contains structured data on individual auto insurance claims, including a mix of anonymised categorical and continuous features.

Although this dataset focuses on auto insurance, which differs from the speciality commercial lines relevant to companies like Accelerant, it will serve as a suitable proxy to develop and test claim severity prediction models. The modelling techniques applied are expected to be transferable across insurance domains.

3.2. Data Cleaning

To ensure data quality, the following steps will be undertaken:

- Handle missing values (if any)
- Verify and correct data types
- Remove duplicate records
- Check for and address inconsistent or invalid categorical entries
- Confirm all fields contain valid, usable data

3.3. Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Understand feature distributions via histograms, boxplots, and bar charts. Identify outliers using descriptive statistics and visualisations
- **Bivariate Analysis:** Explore variable relationships with scatter plots, correlation, boxplots, and crosstabs
- **Multivariate Analysis:** Discover deeper patterns using pair plots, clustering, or PCA
- **Patterns & Trends***:* Identify clusters, risk profiles, variable combinations and anomalies

3.4. Data Wrangling

Based on EDA insights:

- Select relevant features; drop redundant or highly correlated ones
- Engineer new features from existing data
- Encode categorical variables
- Normalise or standardise numerical data (if needed)
- Apply dimensionality reduction or binning where needed
- Sample or filter data if needed to manage scale

3.5. Modelling

Models explored will include:

- Linear Regression (simple and multiple)
- Polynomial Regression
- Ensemble methods such as Random Forest and XGBoost

---

4. Results

Model performance will be assessed using metrics such as:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Cross-validation and hyperparameter tuning will be applied to enhance model robustness, where appropriate.

Feature importance and explainability methods (e.g., SHAP values or feature importance scores) will help interpret results and demonstrate their relevance to risk assessment and pricing strategies.

---

5. Key Insights

After completing EDA and initial modelling, key insights will be summarised to guide final model selection and interpretation. This includes:

- **Main patterns** in feature distributions and relationships
- **Strong correlations** and significant bivariate trends
- **Outliers or anomalies** that may merit further investigation
- **Feature importance** insights to inform modelling strategy

These findings will help refine the approach, support model explainability, and demonstrate practical relevance to real-world insurance applications.

6. Deliverables

- **Final Trained Model & Evaluation Report:** A comparative evaluation of multiple models, with a final selected model for claim severity prediction, accompanied by performance metrics and validation results.
- **Visual Insights:** Clear, interpretable visualisations highlighting key drivers of claim severity.
- **Dashboard or Jupyter Notebook:** An interactive platform showcasing the analysis, model development process, and findings for transparency and reproducibility.
- **Operational Proposal:** Suggestions for leveraging the model to complement or critically assess Accelerant's current advanced solutions by validating—or otherwise questioning—key drivers and enhancing interpretability for improved risk assessment and model governance.

---

7. Timeline

At 14 hours a week the proposed schedule is:

Week	Goal
1	Data sourcing, cleaning, and exploration
2	Feature engineering and initial model development
3	Model tuning, evaluation, and interpretation
4	Visualisation, documentation, and final reporting