# 📄 Executive Summary – Motor Claims Fast-Track Prediction with GenAI

### 🔗 GitHub Repository

**Source Code:** <https://github.com/Gobinath1994/Motor-Claims-Fast-Track-Prediction-with-GenAI>

## 1. Problem Statement

Motor insurance providers handle thousands of claims every month. While many are complex and require human intervention, a large proportion are routine and low-risk—making them ideal for **fast-track processing**. Fast-tracking such claims significantly reduces operational effort, turnaround time, and improves customer satisfaction.

This project addresses the following business question:

***“Can we predict whether a motor claim is eligible for fast-track processing using structured metadata and unstructured free-text descriptions?”***

## 2. Dataset & Technical Challenges

The dataset provided consists of **10,000 historical claims** with:

* **Structured data**: Vehicle make/model, mileage, accident type, customer tenure, etc.
* **Unstructured data**: Free-text “damage\_description” field containing claims narrative.

### Key Challenges:

* Inconsistent formats in text data (misspellings, informal language).
* Imbalanced target distribution (fewer fast-tracked claims).
* Noisy or missing fields like mileage and textual descriptions.

## 3. Feature Engineering

The feature engineering pipeline included a rich blend of structured transformations and NLP techniques:

### Structured Features:

* Derived vehicle age from year
* Claim-to-tenure ratios
* Binning risk features (e.g., damage severity groups)

### NLP Features (on damage\_description):

* **TF-IDF** (Top 30 words)
* **MiniLM Sentence Embeddings** (384-d vector per row)
* **PCA Reduction** to 5 key dimensions
* **KMeans Clustering** to create semantic clusters

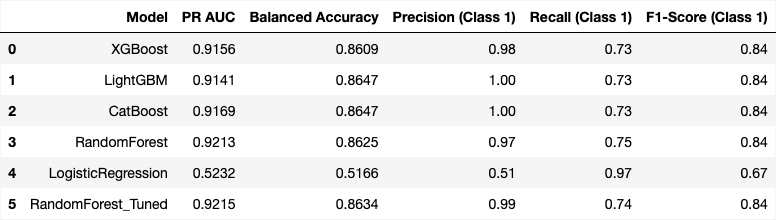
All features were scaled and concatenated for modeling. Feature selection was performed using ensemble voting across 6 different methods (Mutual Info, L1 Logistic, LightGBM importance, Permutation, Boruta, etc.).

## 4. Modeling & Evaluation

Multiple classifiers were trained using Stratified 5-fold Cross-Validation, including:

* **Random Forest**
* **XGBoost**
* **CatBoost**
* **MLPClassifier**
* **LightGBM**
* **LogisticRegression**

### Final Model:



* **RandomForestClassifier (Tuned)**
* **PR AUC**: 0.9156
* **Balanced Accuracy**: 0.8609
* **Threshold optimization** used for maximizing recall of eligible (class 1) cases.

Hyperparameter tuning was performed via Optuna. Class imbalance was handled using **SMOTE oversampling**.

## 5. Model Explainability

**SHAP (SHapley Additive exPlanations)** was used to interpret feature influence:

* Top contributing features:
  + damage\_level\_reported
  + vehicle\_mileage
  + days\_between\_accident\_and\_claim
  + text\_cluster and text\_pca\_1-5

This interpretability was embedded directly into the dashboard for each prediction.

## 6. GenAI Integration – Amazon Bedrock

To enhance downstream processing, the project integrates **Generative AI (Mistral 7B via Amazon Bedrock)** to auto-generate:

* Executive Summaries per claim
* Risk Tags
* Suggested Next Steps

These allow human claim agents to make quicker triage decisions and improve overall process efficiency.

## 7. Deployment & Streamlit Dashboard

The complete solution is deployed on **AWS EC2**, and includes:

* CSV batch upload for bulk scoring
* Live prediction with visual indicators
* SHAP explainability per claim
* Amazon Bedrock-powered LLM outputs
* Results saved to AWS RDS (motor\_claims\_predictions)

**Access**: Dashboard runs via streamlit run and is exposed on port 8501 via EC2 public IP.

## 8. Next Steps & Recommendations

* **Deploy as API** for real-time scoring during claims intake
* **Feedback loop** for user corrections to retrain models
* **Improve LLM prompts** to reduce token cost and refine summaries
* **Drift detection** with rolling monthly claims data

## ✅ Business Impact

* Automated identification of fast-track claims enables near-instant decisions, reducing processing time from days to minutes.
* Operational savings through automation of low-risk claims
* Enhanced human productivity through LLM-assisted reasoning
* Reliable decision-making supported by SHAP explanations

## 👤 Author

Built by **Gobinath Subramani**

A data science and GenAI-driven solution designed for motor insurance modernization.