**📑 Project Summary — InsureScore**

**1. Client Requirements (Demand)**

The client wanted a system that:

1. **Predicts claim probability** → whether a customer is likely to file an insurance claim.
2. **Risk Scoring System** → assign scores (0–100) and segment into **Low, Medium, High risk**.
3. **Customer Behavior Analysis** → study demographics, vehicle details, claim history.
4. **Model Interpretability** → use SHAP and LIME to make model decisions transparent.
5. **Fraud Flagging Logic** → detect suspicious or high-risk cases for manual review.
6. **Business Impact** → personalize premiums, identify high-risk customers, reduce fraudulent payouts.

**2. Data Used**

We worked with publicly available **insurance customer datasets (Kaggle)** that contained information about policyholders and their claim history. The key features were:

* **Customer Demographics**: Age, Gender, Region Code, Driving License status.
* **Policy Details**: Annual Premium, Policy Sales Channel, Vintage (policy tenure).
* **Vehicle Details**: Vehicle Age, Vehicle Damage history.
* **Target Variable**: **Response** (1 = customer filed a claim, 0 = customer did not file a claim).

**3. Steps We Did**

**🔹 Data Preparation**

* Cleaned and explored customer & policyholder data.
* Encoded categorical features (Gender, Vehicle Age, Vehicle Damage).
* Balanced the dataset using **SMOTE** because claims were much fewer than non-claims.

**🔹 Modeling**

* Built **three models**: Logistic Regression, Random Forest, XGBoost.
* Evaluated with **ROC-AUC, Precision, Recall, F1-score**.
* **XGBoost performed the best** and was used as the final model.

**🔹 Risk Scoring System**

* Converted claim probabilities → **Risk Scores (0–100)**.
* Assigned customers into:
  + **Low Risk (0–33)**
  + **Medium Risk (34–66)**
  + **High Risk (67–100)**

**🔹 Model Interpretability**

* **SHAP**: Identified global feature importance.
  + Key drivers: Previously\_Insured, Vehicle\_Damage, Age, Annual\_Premium.
* **LIME**: Explained individual customer predictions in plain terms.

**🔹 Fraud Flagging**

* Rule-based + probability-based logic.
* Example: Customers with **high premium + high predicted risk** were flagged.
* Created a separate fraud-flagged list for review.

**4. Outputs / Deliverables**

1. **CSV Outputs**
   * final\_predictions.csv → risk score, bucket, fraud flag for each customer.
   * fraud\_flagged\_customers.csv → only suspicious customers flagged.
2. **Plots / Visuals**
   * **SHAP Summary Plot** (shap\_summary.png) → global feature importance.
   * **LIME Explanation File** (lime\_explanation\_example.html) → explanation for one customer.
3. **Notebook**
   * Complete Jupyter Notebook (Untitled.ipynb) with data prep, modeling, interpretability, fraud logic.

**5. Results**

* Built an **accurate predictive system** for claim probability.
* Assigned **risk scores** and categorized customers into clear buckets.
* Generated **fraud alerts** for unusual high-risk customers.
* Provided **transparent explanations** with SHAP & LIME for underwriting teams.

**6. Business Value (Impact)**

* **Premium Adjustment:** Underwriters can now price policies fairly based on risk.
* **Fraud Reduction:** Fraud team gets early alerts for suspicious claims.
* **Customer Segmentation:** Marketing teams can target **medium-risk customers** for retention and upselling.
* **Transparency:** Explanations build trust in the model among business stakeholders.

**✅ Final Conclusion**

The **InsureScore project** successfully meets the client’s requirements:

* It predicts claims accurately,
* Scores customers into risk levels,
* Flags possible fraud, and
* Explains predictions in simple, understandable ways.

This solution will help the client **personalize premiums, reduce fraud, and improve overall profitability** in the insurance business.