Insurance Claim Prediction and Severity Analysis

- Sahil Jena

CONTENTS

PROJECT OVERVIEW

This project focuses on predicting insurance claims and analyzing claim severity using machine learning models. The dataset includes various features related to policy details, car specifications, and environmental factors. The project aims to provide insights into claim risks and severity to optimize premium calculations and resource allocation.

Problem Statement

BUSINESS CONTEXT

- Optimize Risk : Enhance risk models to reduce financial losses.
- Fair Premiums : Use predictions to calculate competitive insurance premiums.
- Resource Allocation : Efficiently deploy resources to high-risk areas.

OBJECTIVE

- Predict insurance claim occurrence (Yes/No) using classification techniques.
- Assess claim severity via regression, considering factors like claim cost, vehicle characteristics, and policyholder details.

Optimizing risk, ensuring fairness, and enhancing efficiency.

DATA OVERVIEW:

Features: Key variables include policy_tenure, age_of_car, segment, model, age_of_policyholder, population_density, area_cluster, engine, safety features and ncap_rating, offering insights into policyholder and vehicle characteristics.

HANDLING MISSING DATA

No Missing Values.

OUTLIER DETECTION

Extreme values were identified and addressed to avoid skewing model performance.

FEATURE SCALING

Continuous variables (e.g., age_of_car, population_density) were scaled for consistency and better model accuracy.

CATEGORICAL ENCODING

Categorical features (e.g., policy_type, fuel_type) were converted to numerical formats using techniques like one-hot and label encoding for compatibility with machine learning algorithms.

Policy Risk Score = 0.3 × safety score + 0.2 × area cluster risk + 0.2 × engine risk + 0.2 × density risk + 0.1 × car age risk

FEATURE ENGINEERING

Risk-Based Columns

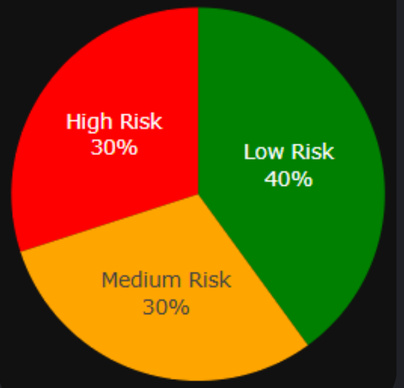
Segment and Area Risk

Safety and Engine Risk

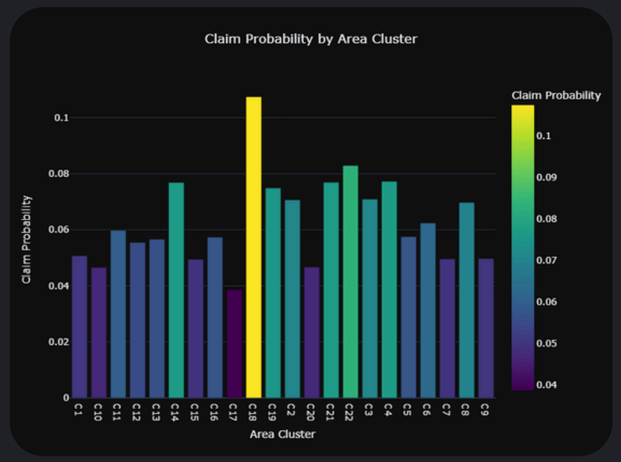
Size and Transmission Risk

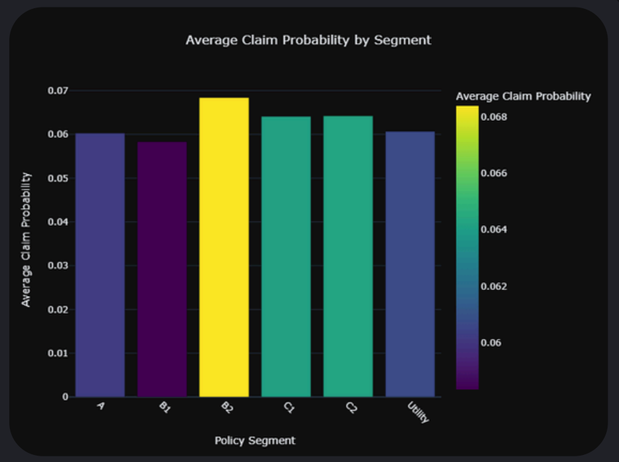
NCAP Risk

- Policy Tenure Risk : policy_tenure was categorized into High Risk, Medium Risk, and Low Risk using predefined bins.
- Car Age Risk : age_of_car was classified into risk levels (Low Risk, Medium Risk, High Risk) based on its age.
- Age Risk : age_of_policyholder was binned into risk levels based on age brackets.
- Density Risk : population_density was segmented into Low Risk, Medium Risk, and High Risk levels.
- Segment Risk : Calculated as the mean claim rate (is_claim) for each car segment.
- Area Cluster Risk : Computed as the average claim rate for each area_cluster.
- Safety Score : Derived as the sum of safety-related binary features such as airbags, ESC, and parking sensors.
- Engine Risk = (max_power + max_torque) / Displacement
- Size Risk = (length x width x height)
- Transmission Risk : Assigned a binary value (1 for Automatic, 0 for Manual).
- NCAP Risk : Calculated as the reciprocal of ncap_rating, indicating higher risk for lower rating.

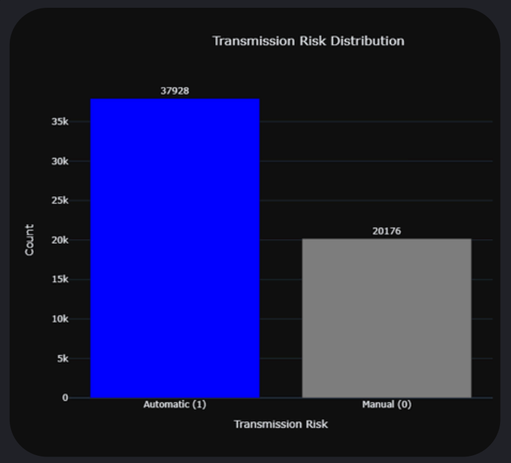


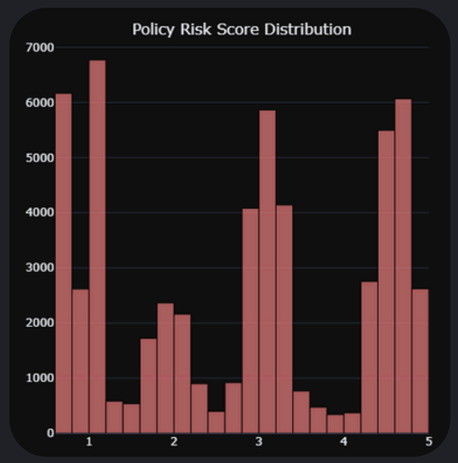
Most policyholders are in low-risk categories for tenure, car age, and population density, offering opportunities for cost-effective policies.





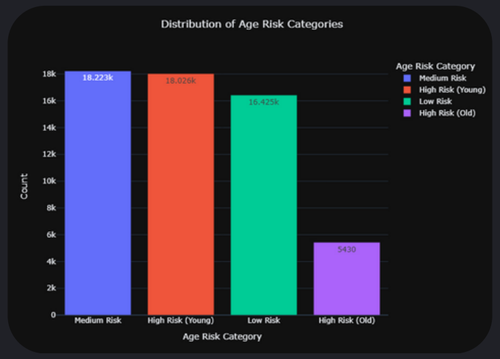
Claim probabilities vary significantly across segments and area clusters, with clusters like C18 and C22 & segments like B2 showing higher risks, highlighting the need for localized strategies.



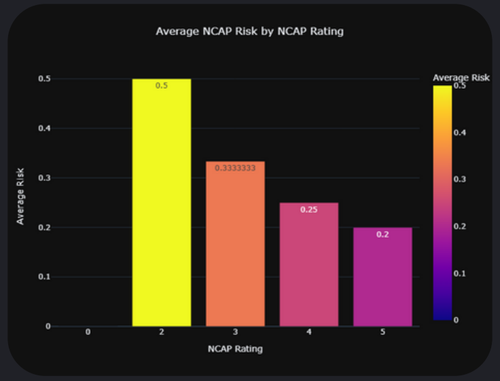


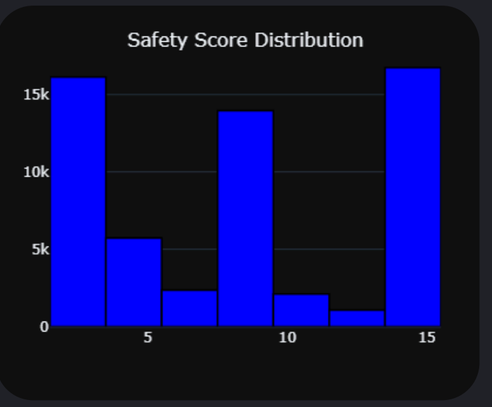
The popularity of automatic vehicles may influence future risk models. Policy risk scores suggest opportunities for tailored premium structures across risk categories.

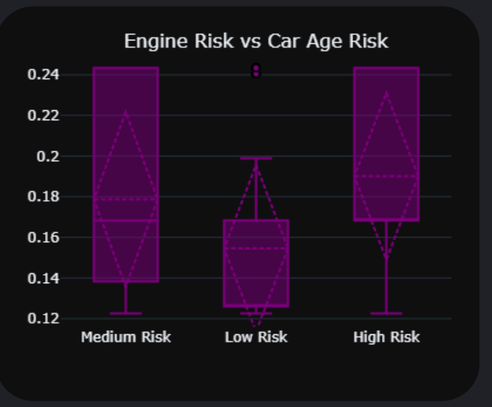
Insights into high-risk groups (e.g., young/old policyholders, high-risk clusters) enable targeted interventions to reduce claims and optimize resources.



Higher NCAP ratings correspond to lower risks, reaffirming the importance of safety certifications.





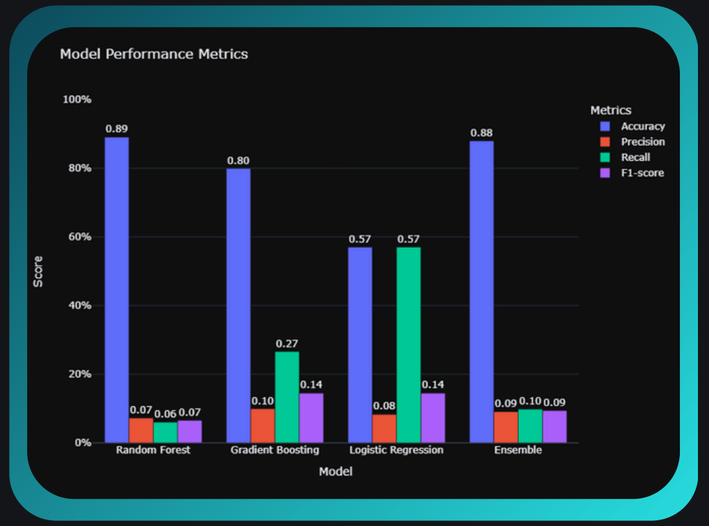


Higher safety scores correlate with lower risks, emphasizing the value of safety features. Engine risk varies with car age, aiding vehicle-specific risk assessment.

Exploratory Data Analysis

Model Comparison for Classification

Balance the Class of training data using SMOTE Technique



Gradient Boosting provided the best balance between precision and recall, especially for the minority class, making it the most effective choice for predicting claim occurrences. The model also supported detailed feature importance analysis, which is crucial for interpretability.

RANDOM FOREST

- Accuracy: 89.06%
- Precision: 0.07%
- Recall: 0.06%
- F1-Score: 0.07%
- Precision Issues with the minority class.

GRADIENT BOOSTING

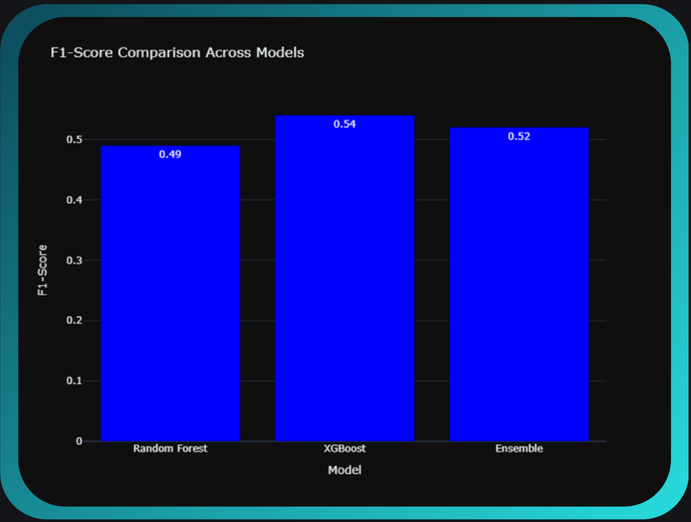
- Accuracy: 79.96%
- Precision: 0.10%
- Recall: 0.27%
- F1-Score: 0.14%
- Best Balance of precision and recall for minority class.

LOGISTIC REGRESSION

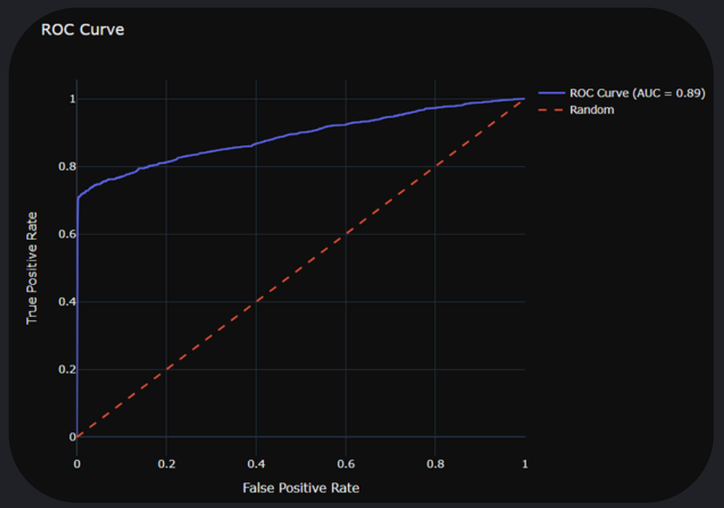
- Accuracy: 57.00%
- Precision: 0.08%
- Recall: 0.57%
- F1-Score: 0.14%
- Struggled with imbalanced data.

ENSEMBLE METHOD

- Accuracy: 87.97%
- Precision: 0.09%
- Recall: 0.10%
- F1-Score: 0.09%
- Similar to Random Forest in performance.



Gradient Boosting Model Performance (Classification)



- This project developed a predictive model to assess claim probability for car insurance, achieving an overall accuracy of 97%. .

The model demonstrates:

- Balanced Performance: Precision (80%) and recall (72%) for claims effectively handle imbalanced data.
- Business Impact: Accurate predictions minimize false negatives and false positives, optimizing claims processing and reducing costs.
- Showcase Value: Highlights advanced data science techniques, real-world applicability, and business relevance.

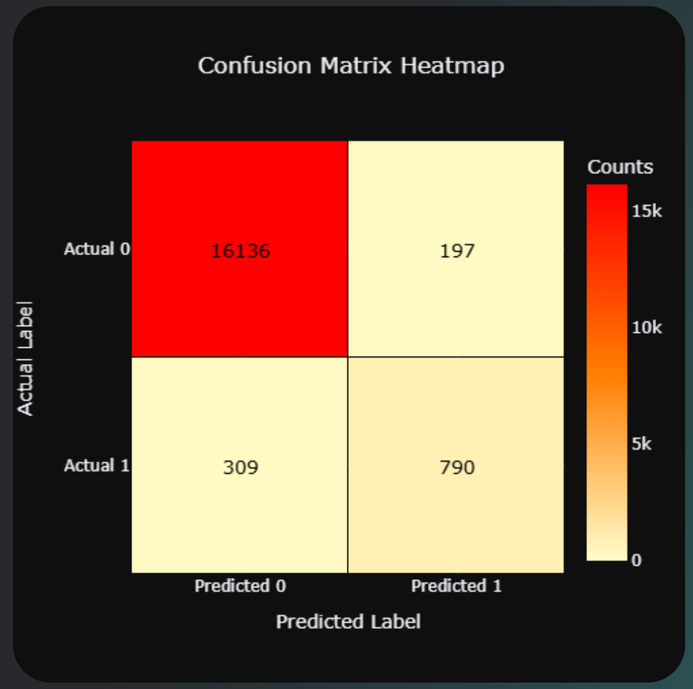
Accuracy: 97%

Precision (Class 1): 80%

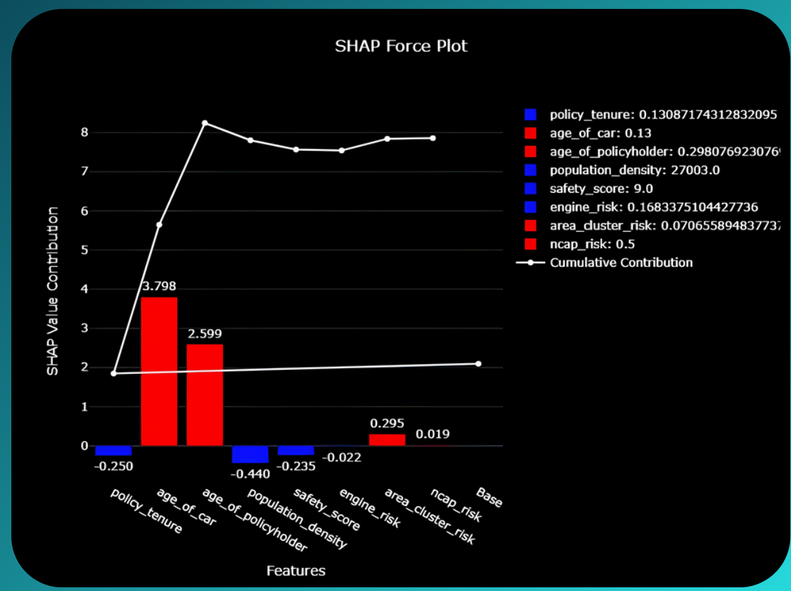
Recall (Class 1): 72%

F1-Score (Class 1): 76%

AUC Score: 89%



Feature Importance Analysis (Classification)

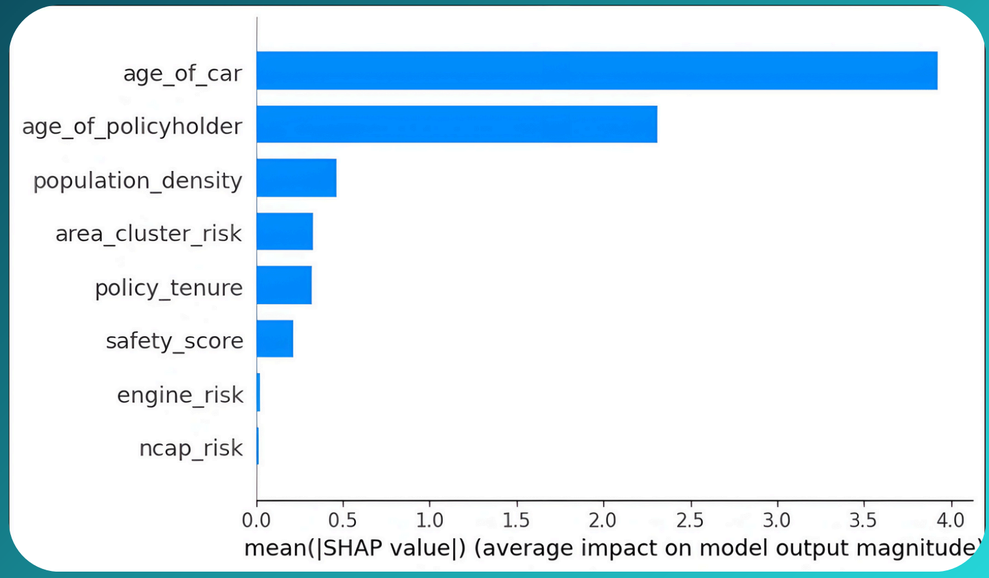


SHAP SUMMARY PLOT

- The Age of Car and Age of Policyholder are the most influential factors, contributing strongly to claim predictions.
- Safety-related features (e.g., Safety Score, Engine Risk) have minor contributions but align with expectations.
- Environmental factors like Population Density and Area Cluster Risk demonstrate their influence on claim predictions.

SHAP FORCE PLOT

- The Population Density and Safety Score exhibit unusually high contributions in this specific instance, deviating from their average trends in the summary plot.
- Age of Policyholder consistently demonstrates a strong positive impact, aligning with the overall trend.



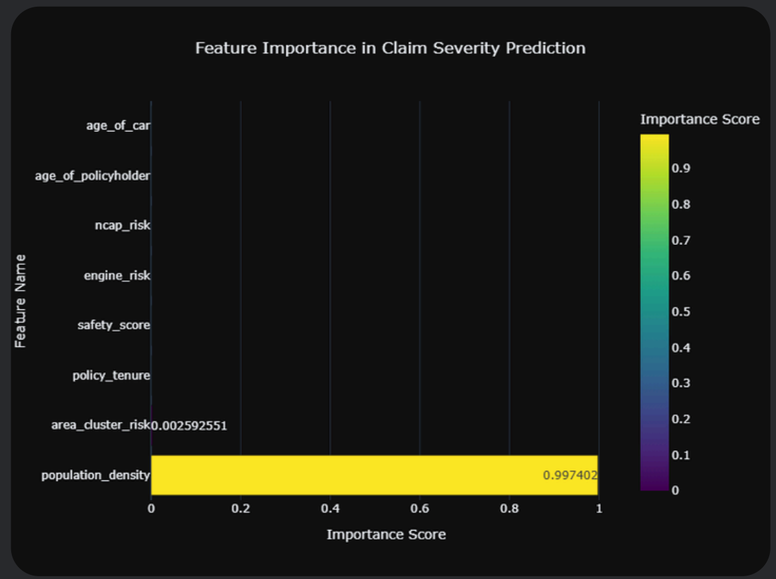
Regression Analysis for Claim Severity

```
df['claim_severity'] = (df['age_of_car'] * 0.1 +
                        df['age_of_policyholder'] * 0.05 +
                        df['population_density'] * 0.2 +
                        df['safety_score'] * -0.2 +
                        df['engine_risk'] * 0.3 +
                        df['area_cluster_risk'] * 0.15 +
                        df['ncap_risk'] * 0.25)
```

- This dataset have not Claim Severity variable. By defining a synthetic target (claim_severity) as a weighted combination of multiple features (like age_of_car, engine_risk, and others), we can understand how these features interact and contribute to the severity of claims.
- This allows us to evaluate whether our model can learn these relationships and predict outcomes that are closely aligned with real-world claim severities.

R^2: 0.9999999911

MSE: 0.1105



- Population Density is the most important factor influencing claim probability, with high importance (0.997).
- Area Cluster Risk and Policy Tenure have minimal impact.
- Engine-related factors (e.g., engine risk, NCAP) show lower importance.

Business Recommendations

Refined Risk Assessment :

- Action: Use the model to identify high-risk policyholders based on features like age_of_car, area_cluster_risk, and engine_risk.
- Impact: Enables accurate risk profiling, reducing unexpected financial losses.

Premium Adjustment Strategies

- Action: Leverage claim probability and severity scores for fair premium pricing:
- High-risk: Higher premiums.
- Low-risk: Discounts/incentives for renewals.
- Impact: Balances profitability and customer satisfaction.

Customized Marketing Campaigns

- Action: Target high-risk regions (e.g., high population_density, specific clusters).
- Impact: Boosts policy sales in underserved areas while mitigating risk.

Enhanced Policy Offerings

- Action: Introduce tailored features for high-risk customers:
- Optional higher deductibles to lower premiums.
- Add-ons like roadside assistance or monitoring systems.
- Impact: Improves retention and attracts price-sensitive customers.

Safety Improvement Initiatives

- Action: Collaborate with local authorities/customers for safety campaigns and reward safe driving.
- Impact: Reduces claim frequency, enhancing brand image and profitability.

Fraud Detection Mechanisms

- Action: Flag anomalies using claim probability data (e.g., high population_density with low engine_risk).
- Impact: Minimizes financial losses by preventing fraud.

Vehicle Feature-Based Incentives

- Action: Offer discounts for vehicles with safety features like ESC, parking cameras, or brake assist.
- Impact: Encourages safer car adoption, reducing overall claims.

Portfolio Optimization

- Action: Analyze claim patterns to identify profitable customer segments.
- Impact: Focus acquisition on profitable segments while managing high-risk exposure.

Continuous Monitoring

- Action: Regularly track model predictions and outcomes to improve accuracy.
- Impact: Adapts to market changes, ensuring consistent performance.

Long-Term Strategic Planning

- Action: Address trends like urbanization (population_density) and evolving vehicle technologies.
- Impact: Future-proofs operations, staying ahead of market risks and opportunities.

thank you