CAR INSURANCE CLAIM PREDICTION

Group 02

W.K.H. Hasara-s16210
Tishani Wijekoon-S16379
Chami Sewwandi-s16028
S.Luxan-s16329

HOW CAR INSURANCE IS CHANGING?

“fraudulent claims cost the U.S. insurance industry over $80 billion every year (FBI report)”

Traditional underwriting methods are slow and outdated, leading to overpriced or underpriced policies.

In Sri Lanka, the motor insurance sector is booming because vehicle ownership is rising.

"Use AI to spot fraud, set fairer prices, and assess risks more accurately."

Sanduni pays more than she should, and Kavindu gets away with lower costs

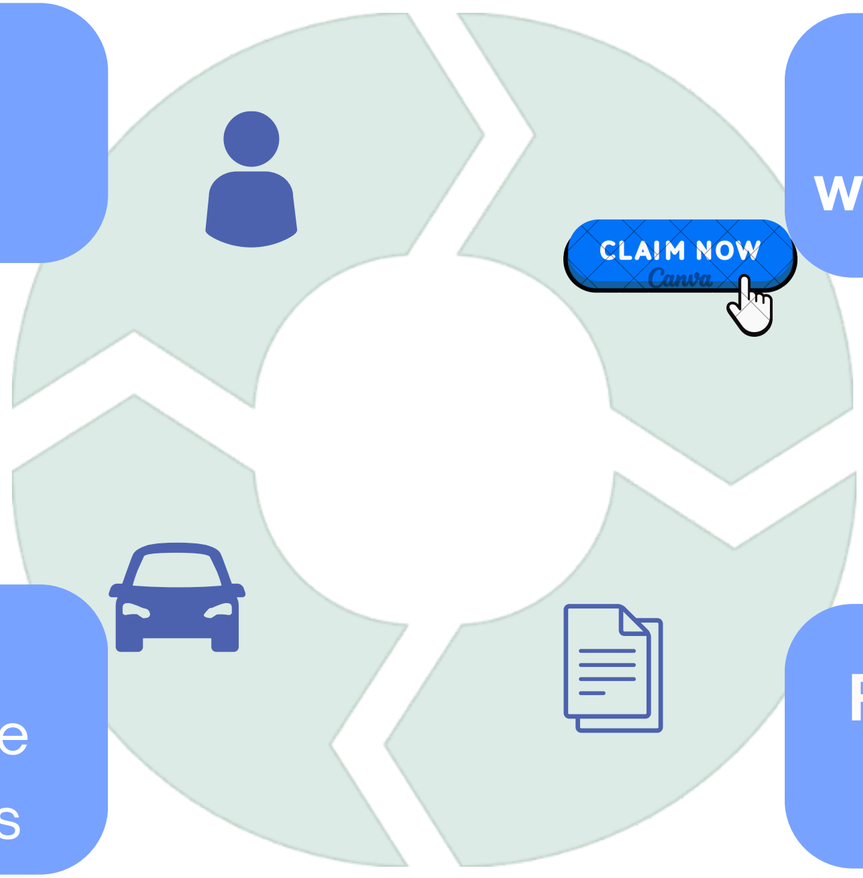
"What if we could predict accidents before they happen?"

DATA SET

44 variables | **58,592 observations**

Policyholder demographics
Age, location

Claim or not within 6 months



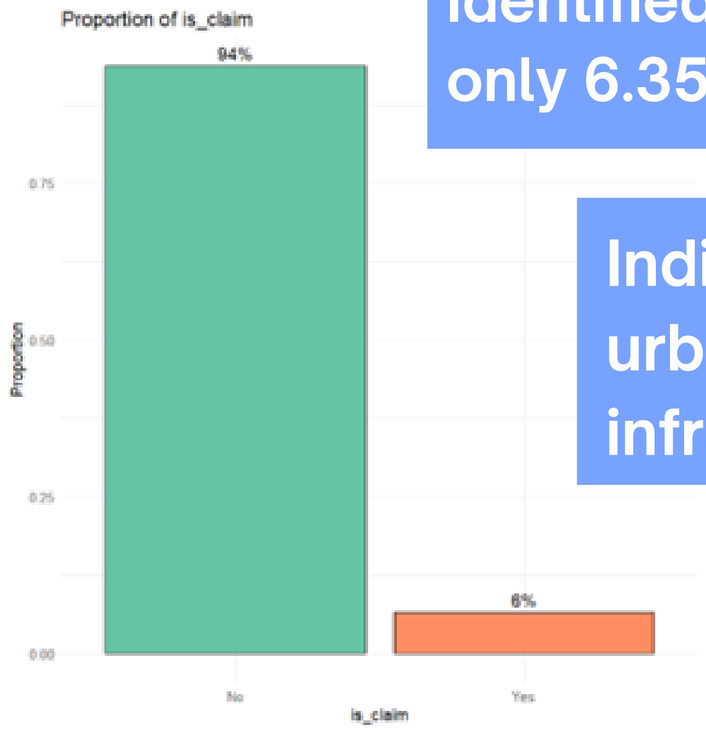
Vehicle attributes
Car model, age, engine power, safety features

Policy details
Tenure

"WHAT DOES THE DATA TELL US?"

Removed 38 variables due to redundancy.

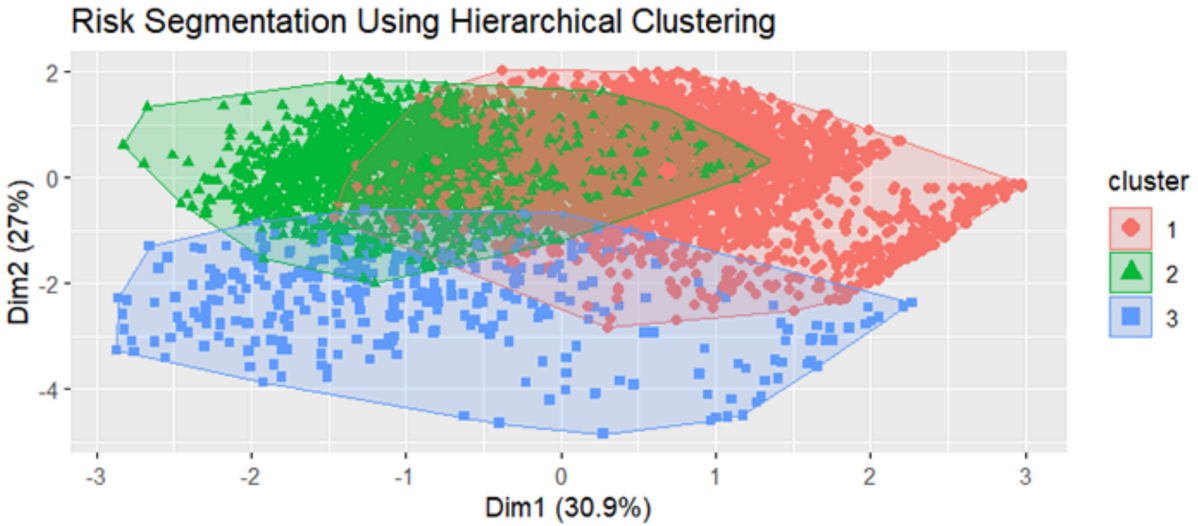
Identified predicting insurance claims as class imbalance, with only 6.35% of policyholders filing claims.



Indicated slightly higher claim rates in low-density urban clusters, possibly due to differences in road infrastructure and accident frequency.

Policy tenure showed a strong association with claims.

"WHAT DOES THE DATA TELL US?"



- Revealed overlapping risk segments.
- Some policyholders didn’t clearly fit into high or low-risk groups, meaning we needed better feature selection.

MODEL FITTING...

MEASURE	LOGISTIC REGRESSION	DECISION TREE	SVC
TRAIN ACCURACY	58.49%	72.81%	62.28%
TEST ACCURACY	55.19%	68.07%	54.28%
PRECISION	95.55%	9.15%	96.09%
RECALL	54.57%	43.13%	53.22%
F1 SCORE	69.47%	15.11%	68.50%

Resampling Technique:

- SMOTE+Undersampling

Results:

- Low Accuracy
- Overfitting
- Poor Recall
- Fails to generalize

Not Ideal For Claim Prediction!

WHY LOW PERFORMANCE?

"Research shows SVM struggles with imbalanced data, favoring the majority class. Even with resampling, performance doesn’t significantly improve."

Farquad, M. A. H., & Bose, I. (2012).

"Decision trees often struggle with predictive accuracy due to their simplicity, but techniques like boosting can significantly improve their performance."

Elements of Statistical Learning

"Logistic regression struggles with imbalanced data, as shown by low precision (0.1533) and AUPRC (0.1586) even after SMOTE.”

Baran & Rola (2022)

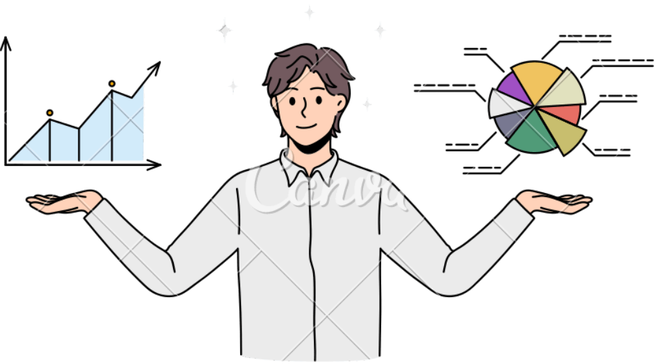
Random Forest vs. XG Boost
The Battle of the Best Predictors

RANDOM FOREST

XG BOOST

- Both fitting using two approaches.

1.SMOTE+ Undersampling
2.Inbuilt class weights



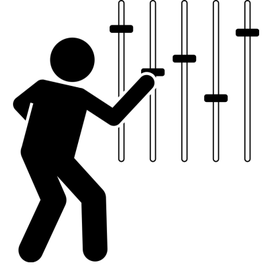
RANDOM FOREST

	CLASS WEIGHTS	SMOTE + UNDERSAMPLING
TRAIN ACCURACY	94.21%	97.23%
TEST ACCURACY	88.05%	82.56%
PRECISION	93.98%	94.08%
RECALL	93.16%	86.75%
F1_SCORE	93.57%	90.27%

overfitting risk

- By using techniques like class weighting and SMOTE+Undersampling, we optimized its performance for our dataset.

XG BOOST



	CLASS WEIGHTS	SMOTE + UNDERSAMPLING
TRAIN ACCURACY	67.02%	93.52%
TEST ACCURACY	63.32%	92.34%
PRECISION	95.10%	93.48%
RECALL	64.03%	98.68%
F1_SCORE	76.53%	96.01%

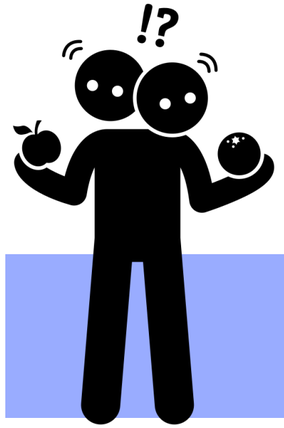
why hyperparameter tuning?

- Boosts model performance & prevents overfitting.
- Finds the best parameter combination efficiently.

why bayesian Optimization?

- Smarter than Grid/Random Search
- Reduces computation time
- Finds optimal values with fewer iterations

- Using **SMOTE + Undersampling** significantly improved accuracy and recall, making it a suitable approach for our dataset.

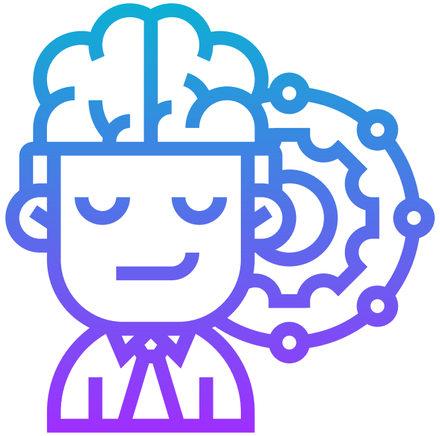


MODEL COMPARISON

	TRAIN ACCURACY	TEST ACCURACY	PRECISION	RECALL	F1_SCORE
LOGISTIC	58.49%	55.19%	95.55%	54.57%	69.47%
SVM	62.28%	54.28%	96.09%	53.22%	68.50%
DECISION TREE	72.81%	68.07%	9.15%	43.13%	15.11%
RANDOM FOREST	94.21%	88.05%	93.98%	93.16%	93.57%
XG BOOST	**93.52%**	**92.34%**	**93.48%**	**98.68%**	**96.01%**

WINNER

XG BOOST



why Boost!

- **Highest Test Accuracy (92.34%)** → Best generalization on unseen data.
- **Highest Recall (98.68%)** → Captures most positive cases, crucial for claims prediction.
- **Best F1-Score (96.01%)** → Balanced performance in precision & recall.

why not other models?

- **Logistic/SVM:** Lower accuracy & recall.
- **Decision Tree:** Moderate performance but lacks robustness.
- **Random Forest:** Good performance but slightly lower test accuracy & recall than XGBoost.

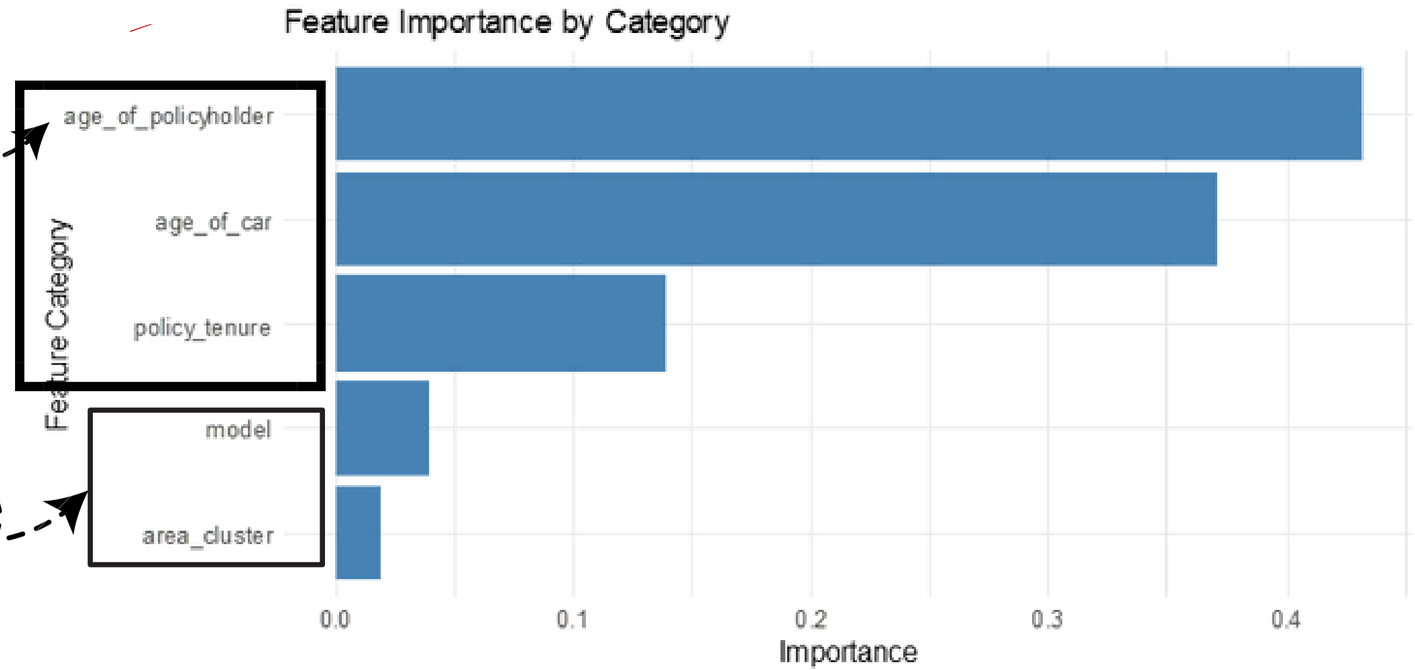
XG BOOST
VARIABLE IMPORTANCE

Most influential factor

AGE OF POLICYHOLDER

Lower impact

MODEL
AREA CLUSTER

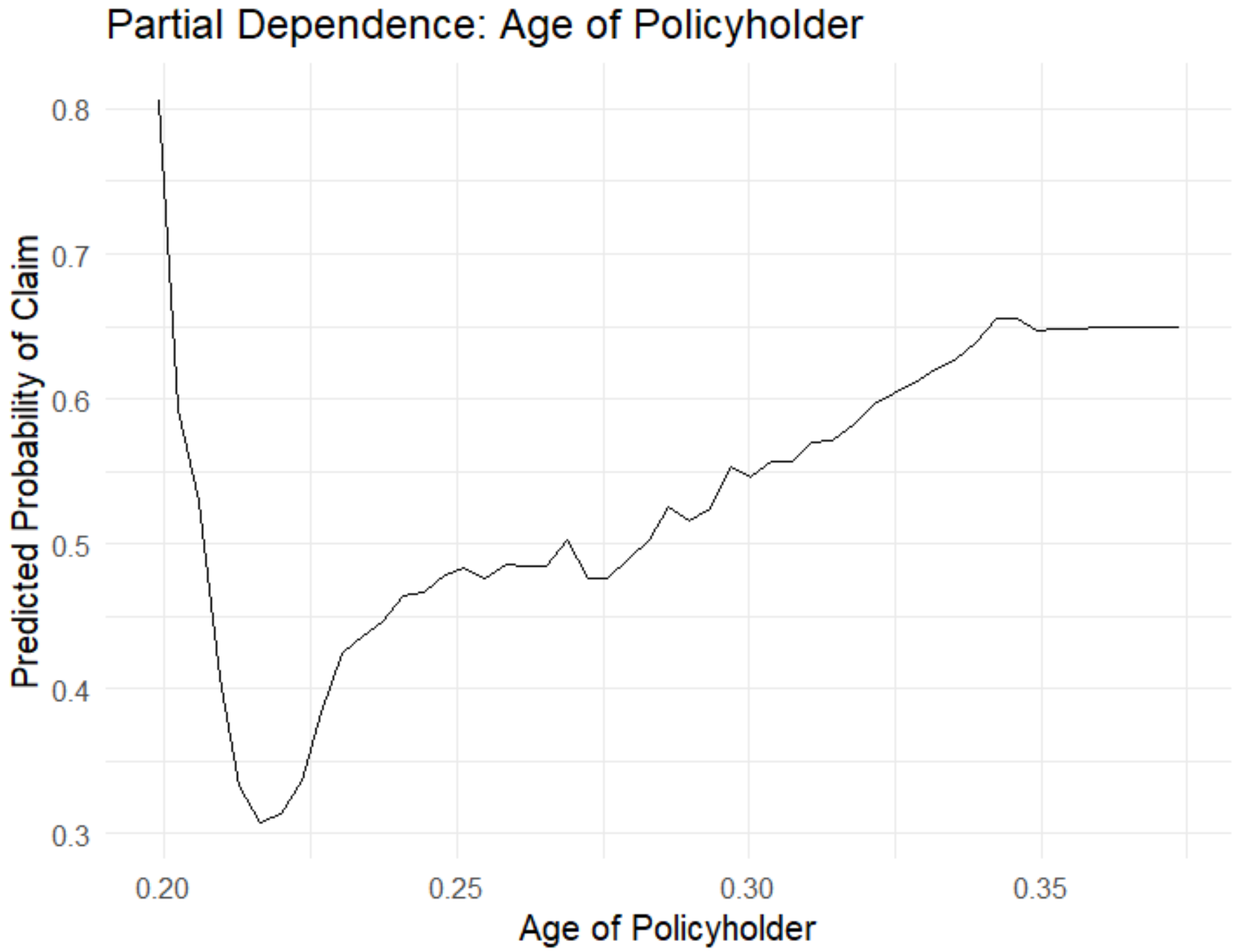


HOW TO EXPLAIN THESE RESULTS?

Partial Dependencies Plots

- Interpret Feature Importance
- Support Decision Making
- Identify Threshold Effects
- Visualize Complex Relationships

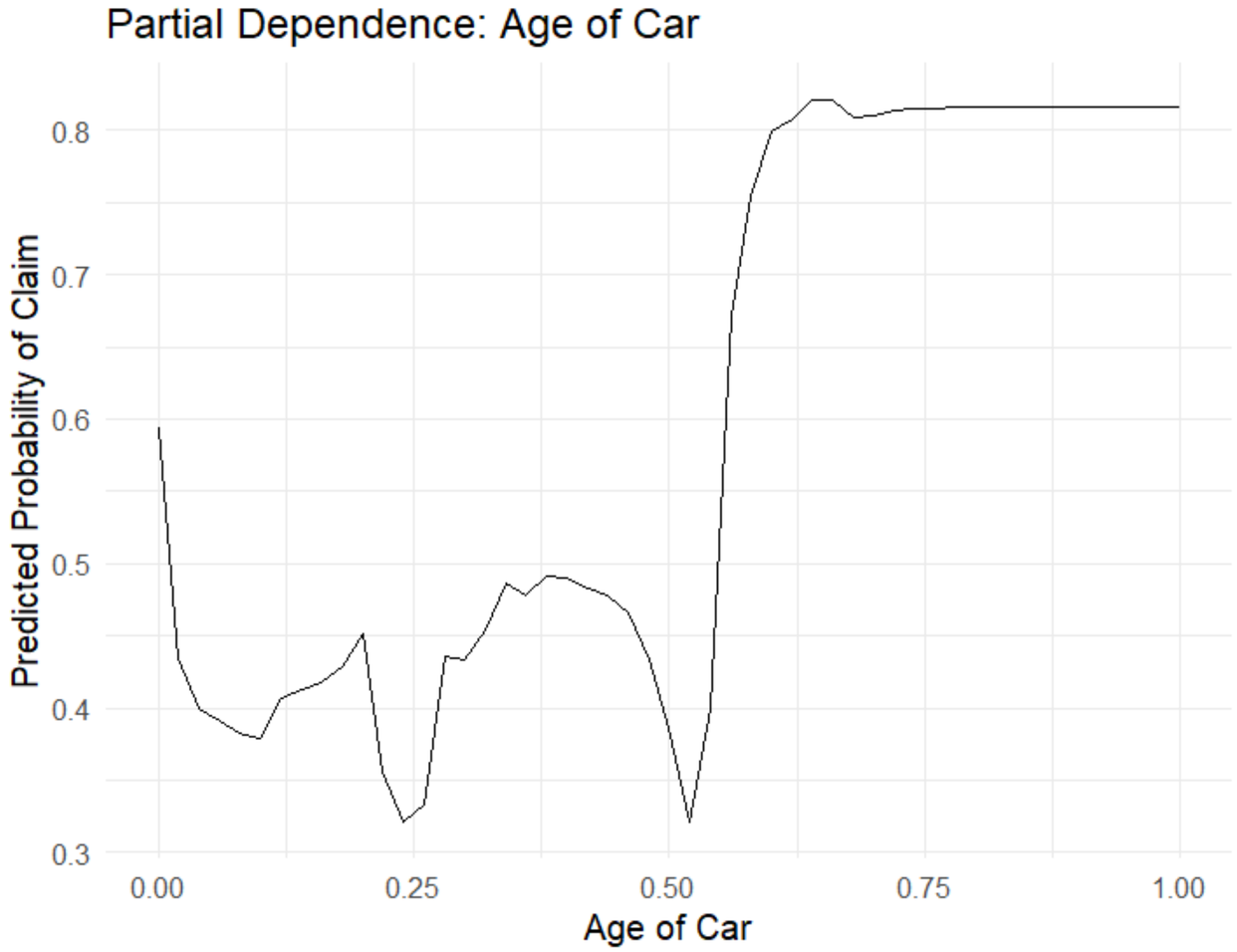
Partial Dependency Plot Of Age of Policyholder



“Drivers under 25 are 26% more likely to be involved in accidents compared to all other age groups.”

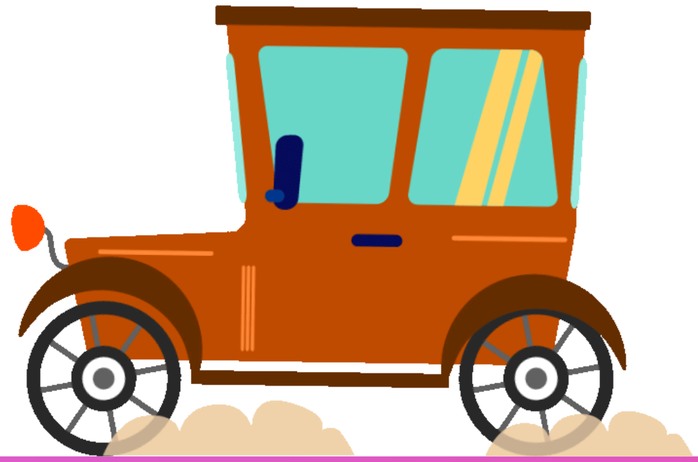
AMI Insurance Statistics(2022)

Partial Dependency Plot Of Age of Car

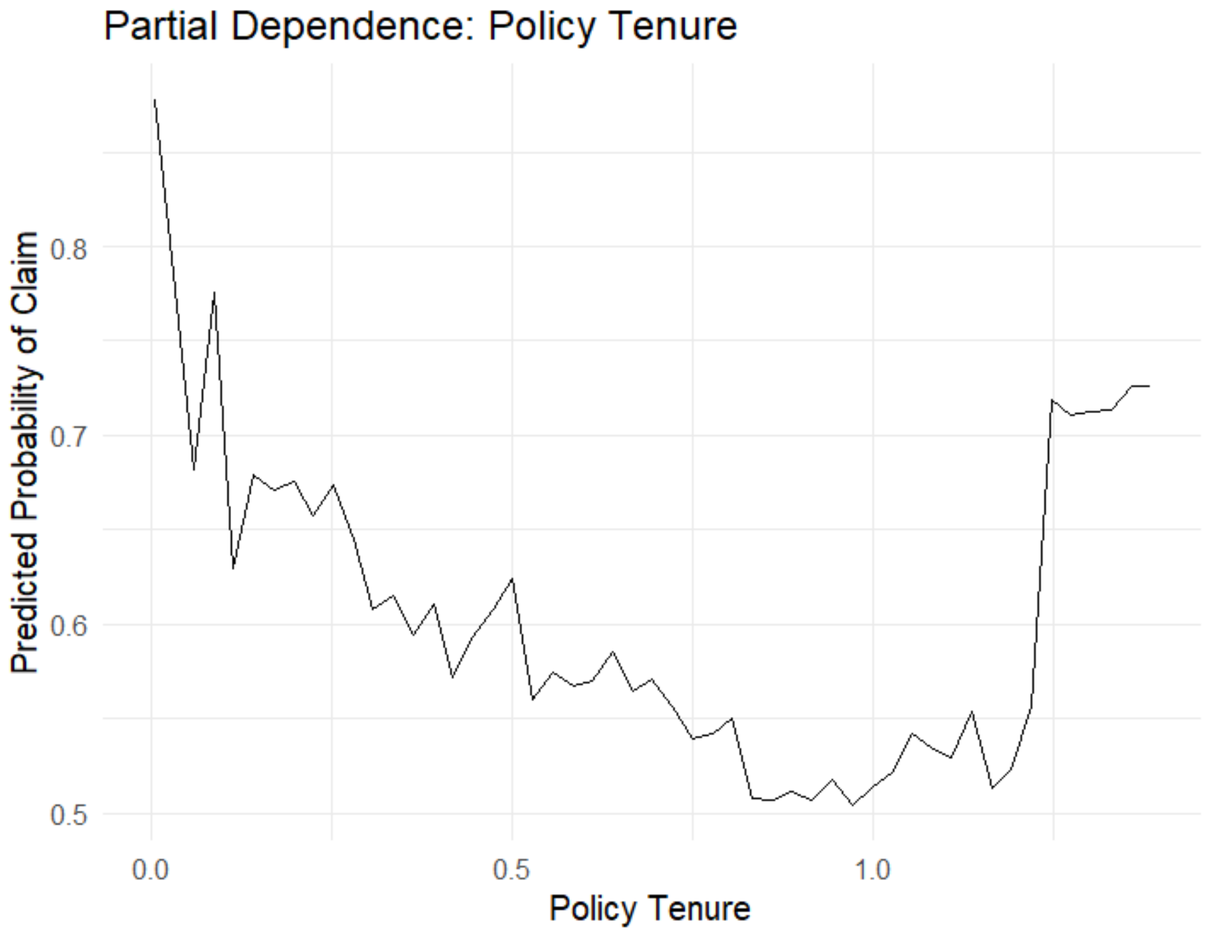


Older Cars

More Claims



Partial Dependency Plot Of Policy Duration



Possible Reasons:

- Initial High Claim Frequency
- Bonus-Hunger Effect

“The Bonus-Malus System rewards policyholders with discounts for staying claim-free, encouraging them to avoid small claims to keep their premiums low. This leads to fewer claims in the mid-term of a policy.”

Wikipedia

Impact of Age of policy Holder

Young drivers

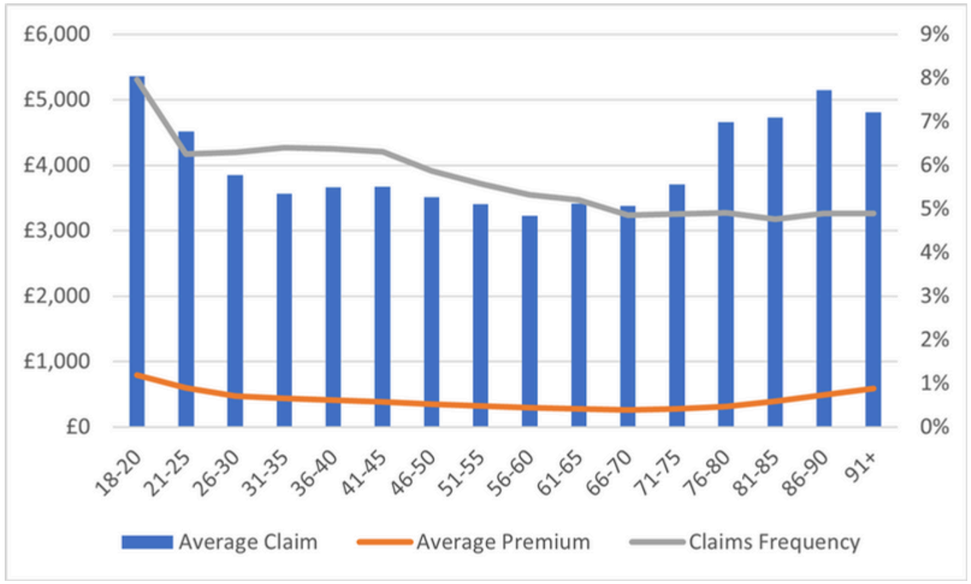
*"Young drivers haven't had the chance to build up experience on the roads. A scary fact is that **1 in 5** young drivers will be involved in a car crash within six months of passing their test."*

-Insurance experiments-

Older drivers

"Experience isn't the answer to everything. Insurers' statistics also show that after the age of 71 drivers become more likely to have accidents and again, they are accidents which are more likely to result in really serious injuries."

-Insurance experiments-



Source: Association of British Insurers (ABI)

1

Road Traffic Injuries

are the leading cause of young people aged 15-29 globally.

Source: Youth for Road Safety

Impact of Age of Car

New Cars Tend to Have Higher Premiums
New cars usually have higher insurance premiums because they are more expensive to repair or replace. Insurers face a greater risk of large payouts if a new car is totaled.

Mid-Life Cars: The Sweet Spot for Insurance Premiums?
Cars that are 3 to 5 years old often have the best balance for insurance premiums. They have depreciated enough to lower replacement costs, but still have modern safety features, reducing the risk of accidents. This makes premiums lower for mid-life cars.

Older Cars May Lead to Lower Premiums (But Not Always)
Older cars often have lower insurance premiums due to their reduced value and lower repair or replacement costs. However, this doesn't always apply universally, as factors like the car's condition or model may still affect premiums.

Source:

Automobile MAGAZINE

"However, it's essential to bear in mind that while the age of your car does impact your insurance premiums, it's just one piece of the puzzle."
-Budget Insurance-

Impact of Policy Tenure

Short-term insurance policies, typically lasting less than a year, may have fewer claims as they often provide limited coverage. In contrast, longer-term policies offer broader protection, increasing the likelihood of claims over time.

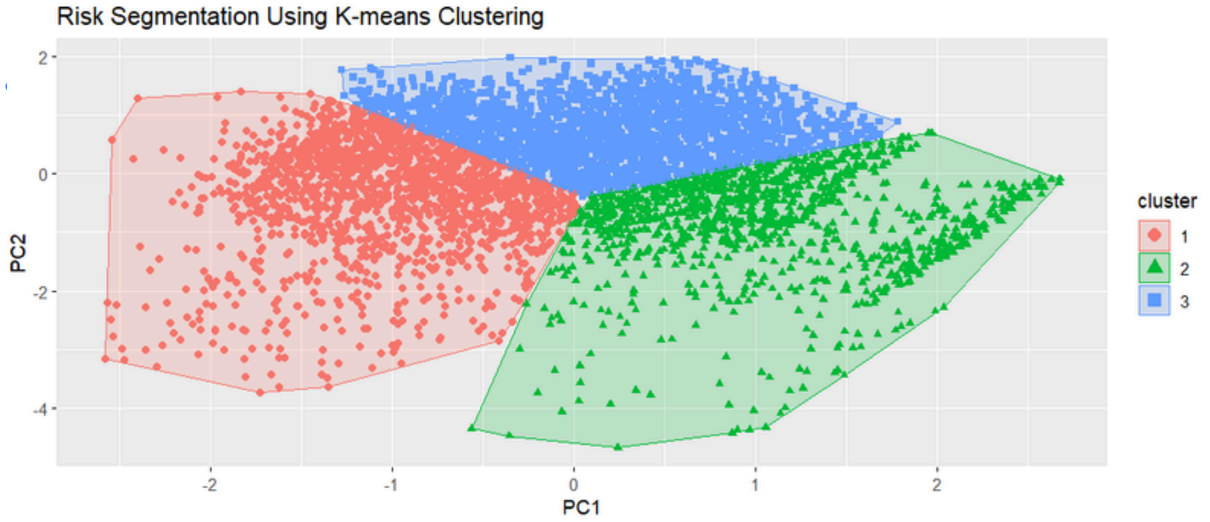
The length of an insurance policy affects its cost. Longer policies usually have higher prices because insurance companies see them as riskier since they last longer.

The type of coverage also affects claims. Comprehensive policies may lead to more claims, while limited coverage policies may have fewer claims, no matter how long they last.

Source: FasterCapital – Policy Duration & Written Premiums

Risk Segmentation Analysis

- The risk segmentation analysis provides a data-driven foundation for insurers to develop customized pricing models and targeted policy adjustments based on policyholder risk levels
- By using the important variables, we obtained the feature importance using XG boost, we applied K means clustering to group the individuals. However, we obtained a low mean Silhouette Score (0.23) for our clustering, but this is a valuable tool for a insurer.



Suggestions for improvement

- Enhance Data Collection & Feature Engineering: Integrate real-time driving behavior (telematics, GPS), policyholder details (employment, credit score, claim history), and external factors (weather, road quality, economic trends) to improve risk assessment.

Conclusion

- the XGBoost model with data balancing techniques performed best, achieving 92.34% accuracy and 98.68% recall. It effectively identifies potential claims while minimizing false positives, providing a reliable tool for insurers' risk assessment.

- Feature importance analysis showed that the Age of the Car, Policy Tenure, and Policyholder Age are the key factors in predicting claims, highlighting the importance of vehicle details and policy length in assessing risk.

Better Risk Management and Policy Optimization:

- Use data to set appropriate prices based on risk levels.
- Apply stricter terms for high-risk customers, closer monitoring for medium-risk, and incentives for low-risk customers.

Future Improvements:

- Incorporate external risk factors.
- Enhance fraud detection methods.
- Leverage real-time data to improve accuracy.

REFERENCES

- 5 Techniques to Handle Imbalanced Data For a Classification Problem. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to handle-imbalanced-data-for-a-classification-problem/
- Baran, S. a. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. arXiv preprint arXiv:2204.06109.
- Decision Tree Pruning. (n.d.). Retrieved from KDnuggets: https://www.kdnuggets.com/2022/09/decision-tree-pruning-hows-whys.html
- Sahai, R. a.-A.-H.-S. (2022). Insurance risk prediction using machine learning. In The international conference on data science and emerging technologies (pp. 419--433).
- SMOTE for Imbalanced Classification with Python. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using smote-techniques/
- Support Vector Machine (SVM) Algorithm. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a complete-guide-for-beginners/
- Trees and Classification. (n.d.). Retrieved from https://fderyckel.github.io/machinelearningwithr/trees-and-classification.html
- XGBoost. (n.d.). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/xgboost/
- Imbalanced Predictions by Stella Safstrom https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9089954&fileOId=9089955

Thank You