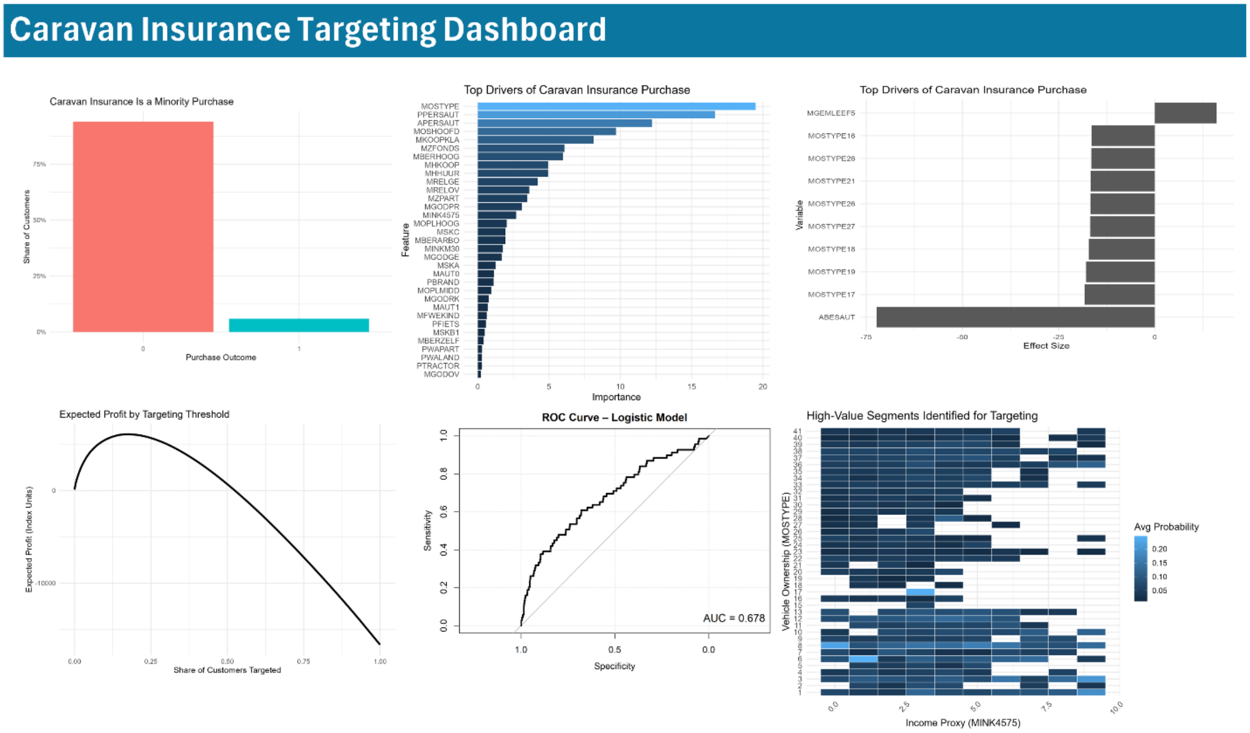
Caravan Insurance Buyer Analysis (COIL 2000)



1. Problem

The decision at stake in this analysis is how an insurance firm should allocate limited marketing resources when promoting a niche product with low natural demand. Caravan insurance is purchased by only a small fraction of customers, which means that untargeted marketing produces low conversion rates and high waste. The strategic objective is therefore to identify a subset of customers who are substantially more likely than average to purchase the product. The model must support a ranking decision: whom to contact first, and how much efficiency can be gained relative to random outreach. In practical terms, the value of the analysis is measured by lift and precision within the highest-scoring segments.

2. Data Reality

The COIL 2000 dataset reflects the realities of operational insurance data. It consists of 86 variables combining product ownership indicators and socio-demographic attributes derived from ZIP-code level information. Although the dataset contains no formal missing values, it suffers from informational sparsity: many variables exhibit low variance, weak individual predictive power, or serve as indirect proxies rather than direct measures of customer intent.

A central constraint is the severe class imbalance. Only about 5–6 percent of customers in the dataset have purchased caravan insurance. This imbalance makes naïve accuracy metrics misleading and increases the risk of models that appear strong statistically but add little business value. In addition, socio-demographic variables are aggregated at the area level, introducing ecological bias and limiting the interpretability of coefficients. Finally, the data is cross-sectional and static. There is no temporal ordering, behavioral history, or feedback loop, which constrains both causal interpretation and long-term generalization.

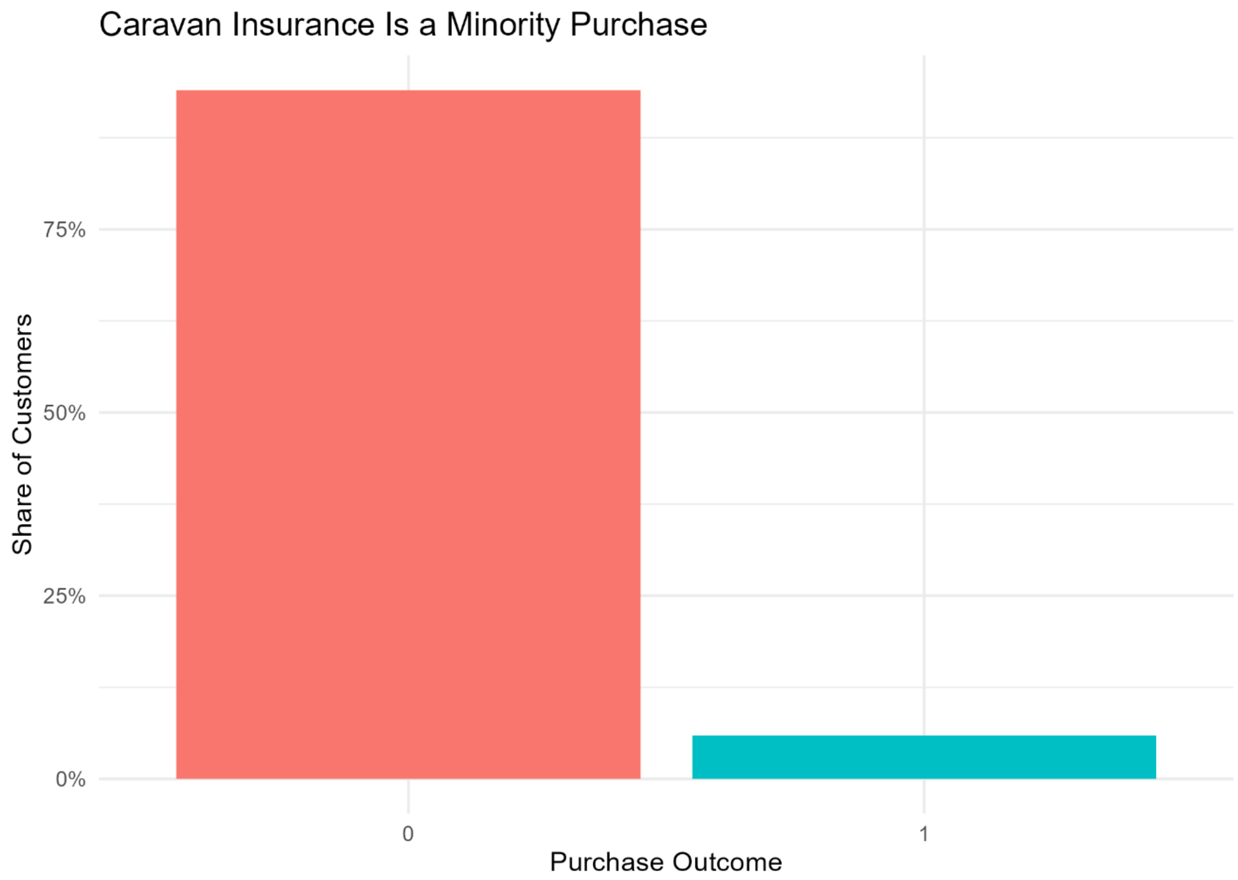
3. Method

Given these constraints, the modeling approach prioritizes robustness, interpretability, and ranking performance over algorithmic complexity. Logistic regression serves as the baseline because it provides a transparent benchmark and establishes whether any signal exists beyond random noise. To address multicollinearity and reduce variance across the large number of correlated predictors, regularized logistic regression is employed. Penalization helps suppress weak or redundant variables while preserving the strongest signals, improving out-of-sample stability.

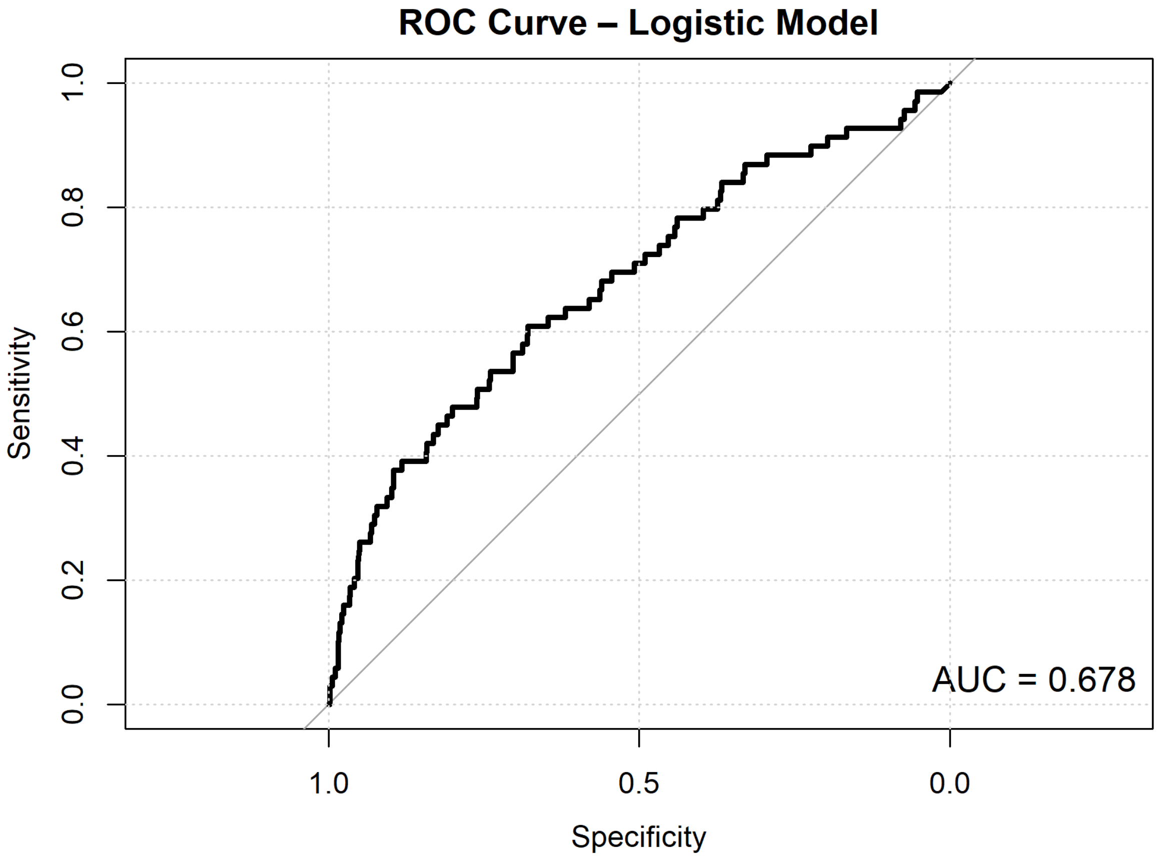
Tree-based models are considered for their ability to capture non-linear interactions, particularly between product ownership patterns and demographic indicators. However, given the small positive class and the risk of overfitting, these models are treated primarily as exploratory tools. The final emphasis remains on regularized linear models because they strike a balance between predictive performance, operational simplicity, and the need for explanation in a business context.

4. Results

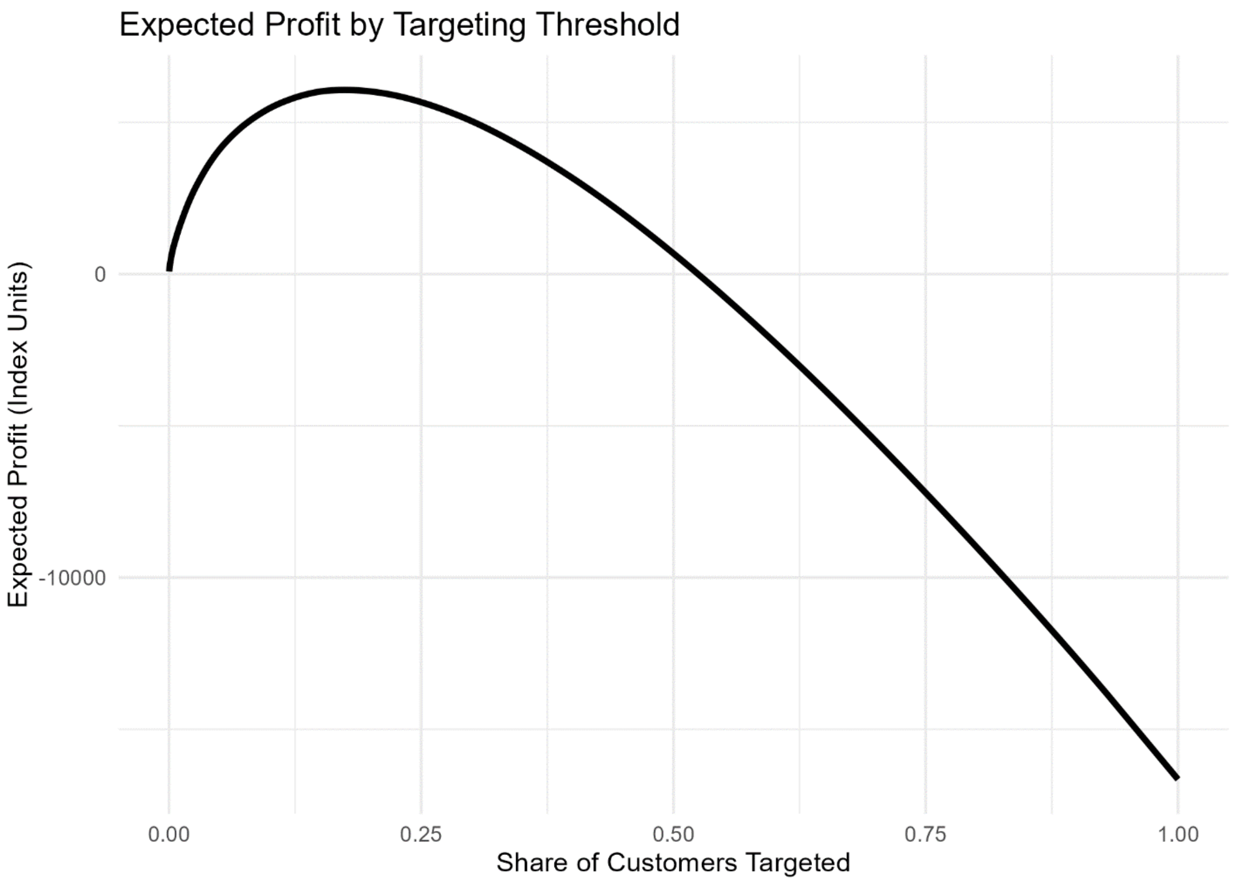
The results confirm that caravan insurance is an intrinsically low-incidence product and that any value from modeling must come from concentration rather than broad prediction. As shown in the class distribution, fewer than 6 percent of customers in the dataset have purchased caravan insurance. This establishes a hard baseline: without targeting, more than 94 percent of outreach effort is expected to be wasted. Any improvement must therefore be judged relative to this structural constraint.



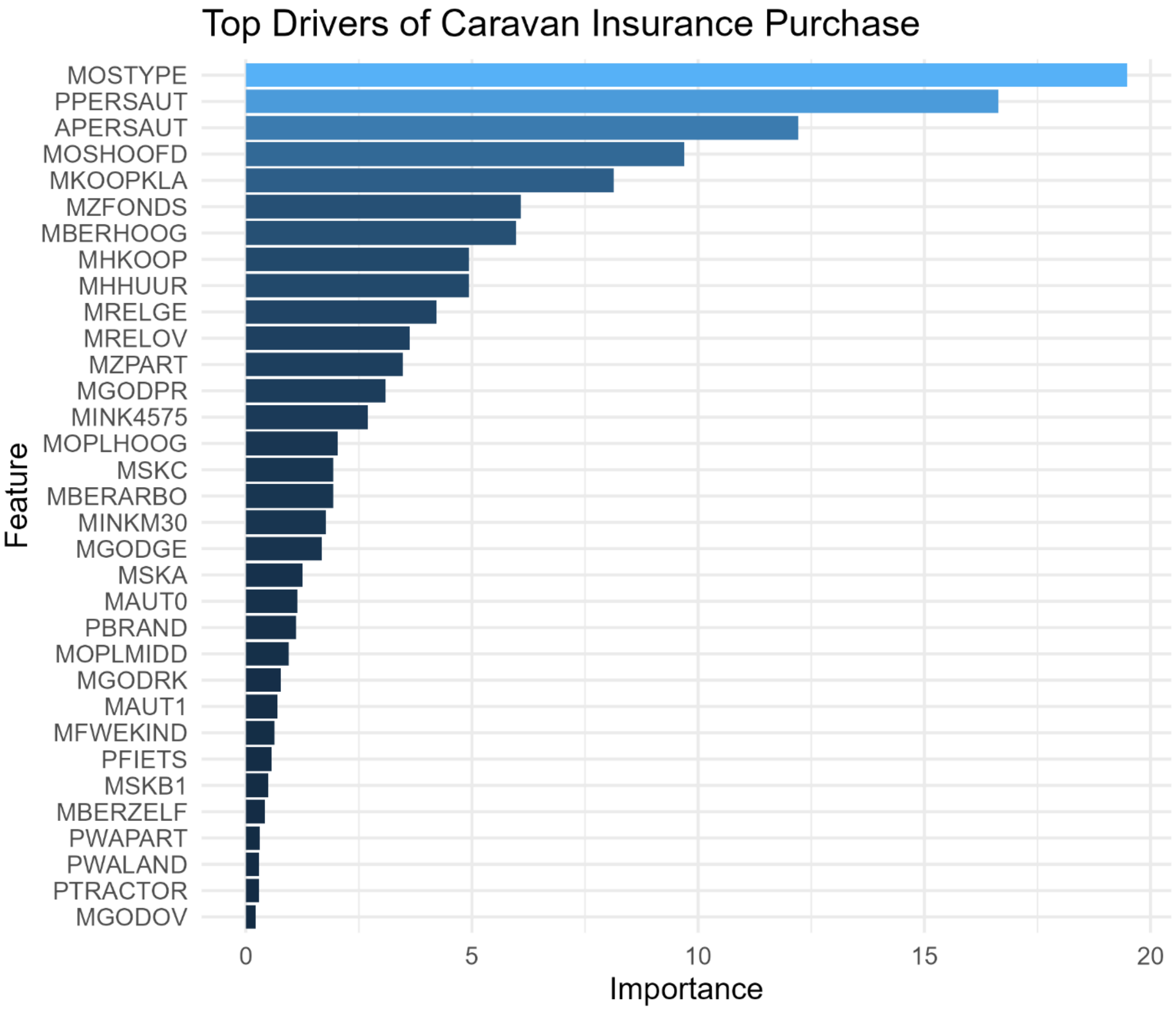
The logistic model demonstrates moderate but meaningful discriminatory power. The ROC curve yields an AUC of approximately 0.68, indicating that the model ranks a randomly chosen buyer above a randomly chosen non-buyer about two-thirds of the time. In isolation this would be unremarkable, but in a heavily imbalanced marketing context it is sufficient to produce economically relevant lift when applied as a ranking mechanism. The ROC curve also shows that gains are concentrated in the early portion of the curve, reinforcing the idea that the model is most valuable at low targeting rates rather than at full population coverage.



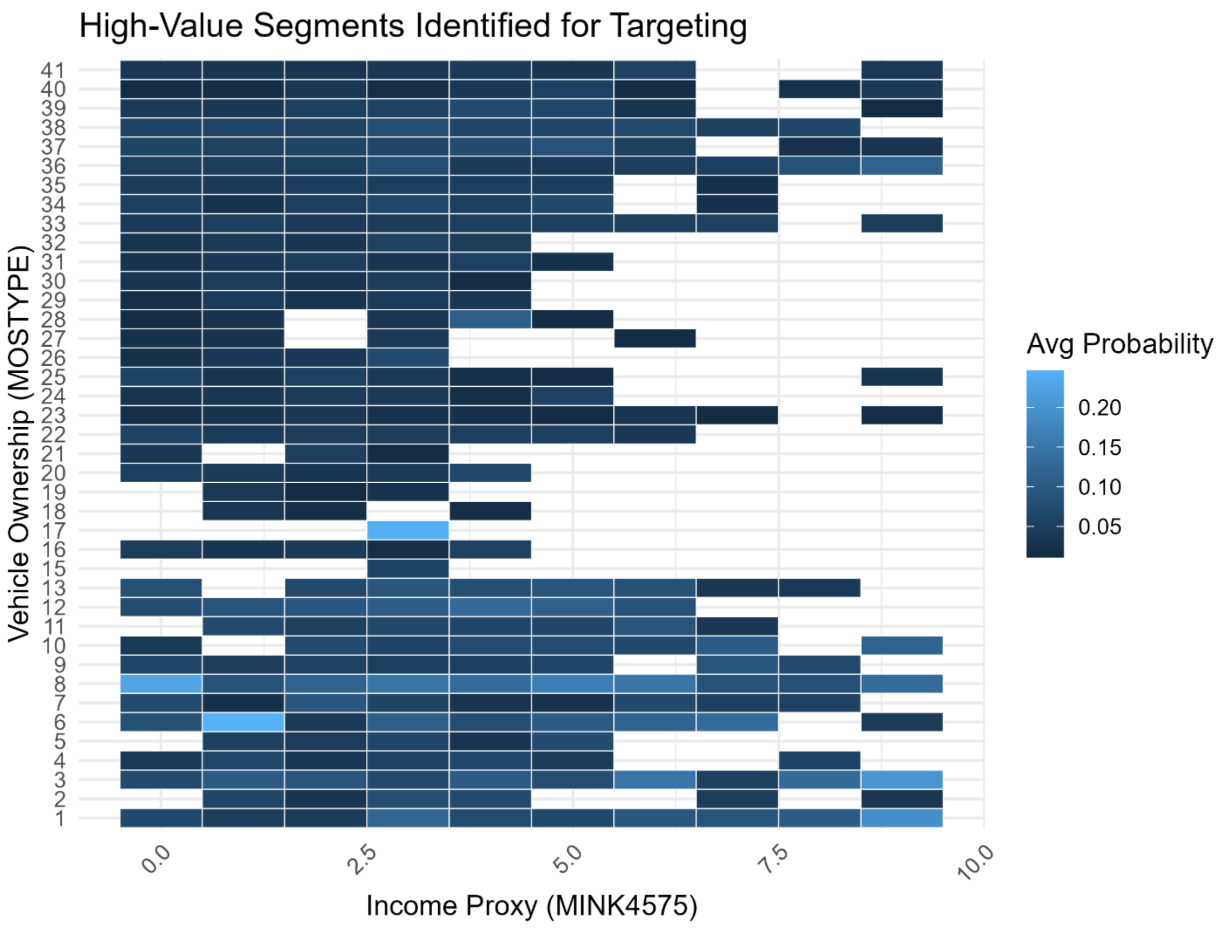
This ranking effect becomes clear when translated into targeting outcomes. The expected profit curve peaks when roughly 15–25 percent of customers are targeted, after which marginal returns turn sharply negative. This implies that beyond this threshold, the declining precision overwhelms the incremental reach. In practical terms, the model supports a disciplined targeting strategy in which only the top-scoring customers are contacted, yielding a substantially higher conversion rate than random selection. Relative to the baseline purchase rate of about 6 percent, the top-ranked segments achieve conversion rates closer to the low-to-mid teens, effectively doubling or tripling marketing efficiency depending on the cutoff chosen.



Feature importance analysis further clarifies where this signal originates. The strongest contributors are product ownership variables, particularly indicators of related insurance products and household insurance portfolios. This aligns with economic intuition: caravan insurance is not a spontaneous purchase but an extension of existing asset ownership and insurance behavior. Socio-demographic variables contribute secondary refinements rather than primary signal. Income proxies and household composition variables matter most when interacting with ownership indicators, not as standalone predictors. No single variable dominates the model; predictive performance arises from the accumulation of many weak signals rather than from any decisive factor.



The segment-level heatmap reinforces this conclusion. Higher predicted purchase probabilities emerge at the intersection of specific vehicle ownership profiles and moderate-to-higher income proxies. Importantly, these patterns are uneven and discontinuous, suggesting that intent is clustered in specific lifestyle configurations. This further supports the use of probabilistic ranking much more than hard rule-based segmentation.



Taken together, the results show that the model does not "solve" caravan insurance prediction in a strong sense, but it materially reshapes the decision space. Instead of treating the customer base as largely homogeneous with rare buyers scattered randomly, the analysis identifies a relatively small, well-defined subset in which purchase likelihood is several times higher than average.

5. Risk

Despite these gains, the results carry several risks that must be managed. First, the observed lift may partially reflect overfitting to historical patterns that do not persist, particularly given the absence of temporal validation. Second, reliance on ZIP-code level socio-demographics introduces instability if population characteristics shift or if the deployment population differs from the training data. Third, focusing on AUC or average performance can mask poor precision at the actual operational cutoff, leading to overconfidence in deployment decisions.

There is also a strategic risk of misuse. Treating the model as a deterministic decision rule rather than a prioritization aid could narrow marketing focus prematurely and reduce learning opportunities. The appropriate interpretation is probabilistic and comparative, not absolute.

To mitigate these risks, the model should be monitored using lift and precision in live campaigns, retrained periodically, and embedded within a broader decision process that allows for exploration and adjustment. Used in this way, the analysis does not eliminate uncertainty, but it converts diffuse uncertainty into structured, economically actionable risk.