Hybrid Non-Heuristic Variable Selection Models via Regularization for Black Box Models Applied to the Insurance Sector

Luciano Galvão
Rafael Moral
lucianogalvao@usp.br
rafael.deandrademoral@mu.ie

Abstract

In this study, machine learning models were tested to predict whether or not a customer of an insurance company would purchase a travel insurance product. For this purpose, secondary data provided by an open-source website that compiles databases from statistical modeling competitions were used. The dataset used presents approximately 2,700 records from an unidentified company in the tourism insurance sector. Initially, the feature engineering stage was carried out, which were selected through regularized models: Ridge, Lasso and Elastic-Net. In this phase, gains were observed not only in relation to dimensionality, but also in the maintenance of interpretative capacity, through the coefficients obtained. After this process, five classification models were evaluated (Random Forests, XGBoost, H2O GBM, LightGBM and CatBoost) separately and in a hybrid way with the previous regularized models, all these stages using the k-fold stratified cross-validation technique. The evaluations were conducted by traditional metrics, including AUC, precision, recall and F1 score. A very competitive hybrid model was obtained using CatBoost combined with Lasso feature selection, achieving an AUC of 0.861 and an F1 score of 0.808. These findings motivate us to present the effectiveness of using hybrid models as a way to obtain high predictive power and maintain the interpretability of the estimation process.

Key Words: Convex optimization; non-linear algorithms; White Box models; Model interpretability; Machine learning.

1 Introduction

Travel insurance is a strategic challenge for the insurance industry, with a significant impact on the customer lifecycle and income generation [1]. Understanding the factors that influence the purchase decision and being able to predict customer behavior are

essential to develop targeted offers and improve business results. In this context, machine learning (ML) methods have shown promise in modeling complex data sets and improving predictive capabilities in the financial services and insurance sectors [2].

On this topic, we aim to demonstrate the use of modern predictive models to ensure predictability in travel insurance contracting, using data made publicly available through an open source modeling competition. The dataset used comprises 2,697 customer records from a company in the tourism insurance sector and includes a mix of demographic, socioeconomic and behavioral attributes resulting in 10 initial attributes.

For this type of product and, more broadly, the insurance market as a whole, there are vast approaches in the literature, from regressive techniques to classification. These seek to explore Machine Learning methods used in predictive studies of travel insurance purchases [3], [4] and [5]. Another focus is on studies that present high dimensionality related to the sector, proposing that the solutions applied to the market be expanded in order to bring broader and more accurate views [6], [7].

Predictive models with complex structures, often called as "black box", tend to have high levels of accuracy, but, on the other hand, offer little transparency. This lack of clarity can be a hindrance, especially in regulatory or strategic contexts or where the results require an understanding of the components of the models. These can be fundamental for taking concrete market actions. The use of five ensemble-based machine learning algorithms is therefore proposed: Random Forest, XGBoost, H2O GBM, LightGBM and CatBoost, associated with regularized models to ensure predictive power and interpretability [8–10].

In this work, we explore non-heuristic statistical approaches as an alternative to increase model interpretability. Penalized regression techniques, such as Lasso, Ridge and Elastic Net, available through the glmnet package [11, 12] in R, provide a solid basis for variable selection, in addition to efficiently dealing with problems such as multi-

collinearity. These techniques contribute to simplifying models, reducing dimensionality and highlighting the most relevant predictors. The result of this type of method are models that combine greater interpretability with a more stable and reliable statistical structure [13–15].

Initially, in 2, a general flow of how the construction of hybrid models is demonstrated, going through the definition of regularized models and their evaluation in 2.1 models in isolation and their results can be observed in 3.1, followed by the blackbox models in 2.2 whose metrics are shown in 3.2. In order to compare the results of the individual and combined methodologies, in 4 there is a discussion about these measures, while in 5 and 5.2 it is demonstrated that the results obtained through simulation are generalizable.

Thus, through this pipeline presented, we attempt to demonstrate the viability of this hybrid approach to reconcile the gains of whitebox and black box models in order to promote new approaches in the insurance sector.

2 Methodology

The initial step involved feature engineering, which expanded the original set of 10 attributes to a total of 35 explanatory variables. The original database contains demographic, behavioral, and socioeconomic information such as: the customer's age, annual income, and total miles accumulated, all of which are continuous. There is also the number of family members, which was discretized in this study. The binary variables: whether or not the customer has a college degree, chronic illnesses, frequency of flying, and whether or not he or she has traveled abroad were treated as such. The type of employment relationship (CLT/Self-employed) and finally the target variable, whether or not the customer has taken out travel insurance.

In addition to the bank's original variables, an expansion was carried out using feature engineering, with the aim of capturing variations in aspects of each client's profile.

New variables were constructed based on proportions, such as the family's per capita income and the ratio between income and age, focusing on the individual's socioeconomic status. Scores were also created, such as the propensity to purchase insurance or not and the perceived risk score, combining factors of income, age, and travel habits into a single measure. The family aspect was considered both through per capita income and by creating indicators of low dependency or families with a larger number of members.

In the professional field, variables on job stability, estimated time of experience in the market and specific employment relationships, such as working in the private or public sector, were included. Patterns directly linked to travel were also mapped through the creation of frequency scores, indicators of international experience and interactions between income and participation in loyalty programs of companies in the sector.

Transformations such as normalization and logarithmic scales were applied to variables such as age, income, miles traveled and chronic diseases, with the aim of reducing distortions and stabilizing variances. Finally, aggregated information based on similar groups was added, such as cluster scores generated by grouping algorithms and moving averages by age groups, enriching the modeling with collective references. Details are available in item **??** of this material.

After this stage of generating the transformed database, a systematic training and validation process was implemented, using stratified data partitioning via k-fold (5 folds). Following the proposed pipeline, the penalized regression models, Lasso, Ridge and Elastic Net, were applied for feature selection, as illustrated in Figure 1.

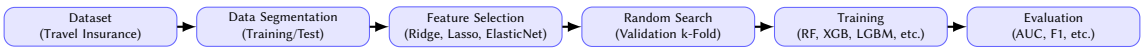


Figure 1: Modeling Flow with Variable Selection, Hyperparametric Tuning, modelling and evaluation

Once the feature selection was done, black-box classification models were trained

using Random Forest, XGBoost, LightGBM, H2O GBM and CatBoost. These models were chosen due to their ability to handle high-dimensional data and capture linear and non-linear relationships. Hyperparameter optimization was performed via Random Search, a computationally efficient alternative to Grid Search, particularly effective in high-dimensional search spaces [16,17]. This combination is defined in this work as hybrid models, i.e., feature selection via regularization followed by optimization and classification with black-box models.

2.1 Regularized Models

Regularized regression models play a central role in contexts characterized by high dimensionality or the presence of multicollinearity between predictors. In this study, the Ridge, LASSO and Elastic Net methods were applied independently in order to evaluate the best approach. These techniques introduce penalization to the regression coefficients to control the model complexity and reduce the risk of overfitting.

Ridge regression, based on the L2 penalty [18], shrinks coefficients towards zero without making them exactly zero, being particularly effective in scenarios with strong correlation between explanatory variables. LASSO regression (L1 penalty) [14], in turn, performs variable selection by forcing some coefficients to be exactly zero.

Elastic Net combines the L1 and L2 penalties through a mixing parameter α [15], offering a hybrid solution. This hybrid approach provides greater flexibility and stability in variable selection, especially in cases involving many correlated features. The pipeline for these models is exposed in Algorithm 1.

Algorithm 1 Training of Regularized Regression Models with Cross-Validation

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$
Output: Sets of selected variables for Ridge, Lasso, and Elastic Net
Split $\mathcal{D}$ into stratified training and testing sets: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$
Set random seed for reproducibility
Standardize predictor variables (z-score normalization)
for model $\in$ {`Ridge`, `Lasso`, `Elastic Net`} **do**
    Set mixing parameter α:
    - $\alpha = 0$ for Ridge
    - $\alpha = 1$ for Lasso
    - $0 < \alpha < 1$ for Elastic Net

    Apply `cv.glmnet` with `family = "binomial"` from the glmnet package [11] for R software:
    - Use `type.measure = "auc"` to select the optimal λ
    - Employ stratified k-fold cross-validation (e.g., $k = 10$)

    Extract coefficients $\hat{\beta}$ from the best model (maximum AUC)
    Select variables where $\hat{\beta}_j \neq 0$
end for
Return: Lists of selected variables for each regularization method

2.2 Integration of Regularized Models and Non-Linear Algorithms

For each of the regularized approaches, models were trained using random samples of the dataset, with control over the random seed to ensure the reproducibility of the results. This step was accompanied by stratified cross-validation to guarantee that the class distribution was preserved across training and validation folds [19]. The tuning process for the hyperparameters λ (penalty strength) and α (in the case of Elastic Net) was carried out based on the maximization of the AUC metric (Area Under the ROC Curve) [20–23].

After selecting the optimal hyperparameters for the regularized models, an analysis

of the estimated coefficients was performed to assess the relative importance of each variable. This analysis enabled an interpretable understanding of the predictors' effects within the travel insurance context, revealing consistent patterns among demographic and behavioral attributes associated with a higher likelihood of product adoption [8–10, 14, 15, 24].

Based on the feature sets selected by the penalized regression models, a second modeling phase was carried out using algorithms with greater predictive capacity. At this stage, black-box models were explored, including Random Forest, Gradient Boosting Machines (GBM), XGBoost, CatBoost, and LightGBM widely recognized for their robustness and superior performance in structured data scenarios [8–10, 25].

These models were tuned through a random search procedure, aiming to identify the best combination of hyperparameters, including the number of trees, maximum tree depth, learning rate and internal regularizations, such as $l2\_leaf\_reg$ in CatBoost that incorporates native regularization mechanisms such as the [25] parameter. This step followed a rigorous experimental protocol, maintaining the previously defined training-testing split and applying cross-validation within the training set for each hyperparameter configuration.

For each black-box model, performance metrics such as AUC, F1-score, precision, and recall were collected to enable a direct comparison of the gains obtained relative to the regularized models. Additionally, variable importance analyses were carried out for the tree-based models using each package's native methods including information gain and permutation-based importance to assess the consistency of key predictors across methods [26, 27]. The entire modeling pipeline is summarized in Algorithm 2.

Algorithm 2 Training of Black-Box Models with Features Selected via Regularization

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$
Preprocessing: Split $\mathcal{D}$ into training set ($\mathcal{D}_{\text{train}}$) and test set ($\mathcal{D}_{\text{test}}$), using a fixed random seed
Step 1: Feature Selection via Regularization
Fit regularized regression models (Ridge, Lasso, or Elastic Net) on $\mathcal{D}_{\text{train}}$ with cross-validation to determine λ^*
Select feature subset $\mathcal{S} = \{x_j : \hat{\beta}_j \neq 0\}$
Step 2: Training Black-Box Models Using $\mathcal{S}$
for model $\in$ {*Random Forest*, *XGBoost*, *CatBoost*, *H2O GBM*, *LightGBM*} **do**
    Define hyperparameter search space and draw:
    - Number of trees: $T \sim \text{Uniform}(100, 1000)$
    - Maximum depth: $d \sim \text{Uniform}(3, 15)$
    - Learning rate: $\eta \sim \text{LogUniform}(0.001, 0.2)$
    - Model-specific parameters (e.g., `l2_leaf_reg` for CatBoost)

    Random Search with k-fold Cross-Validation:
    for each randomly sampled 5 times configuration h **do**
        Evaluate mean AUC across validation folds on $\mathcal{D}_{\text{train}}$
    end for
    Select optimal hyperparameters h^*
    Train final model using h^* on $\mathcal{D}_{\text{train}}$
    Evaluate test performance on $\mathcal{D}_{\text{test}}$: AUC, F1-score, Precision, Recall
    Estimate feature importance (e.g., information gain, Gini index, permutation)
end for
Output: Final trained models using $\mathcal{S}$, optimal hyperparameters h^*, test metrics, and variable importance rankings

3 Results

Subsection 3.1 presents the results obtained from the application of penalized regression models Ridge, Lasso, and Elastic Net in the task of predicting a binary response variable. These models were selected for their ability to perform variable selection or shrinkage through coefficient penalization, which helps to mitigate overfitting while also identifying the covariates with the greatest contribution to the response. In the visual

presentation, it was decided to make a cut of the 10 most relevant variables considering the range of coefficients, seen in 2, while also presenting comments on the interpretability of the coefficients.

3.1 Regularized Models

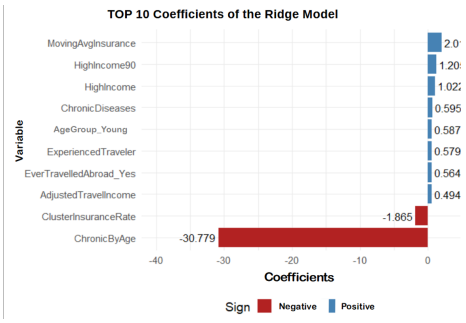
The analysis of the regularized models was performed on three main axes: the relative importance of the variables, the trajectory of the AUC metric as a function of the penalty parameter λ (via cross-validation) and the confusion matrices derived from the performance on the test set.

Figure 2 shows the top 10 coefficients estimated by the methods. The regularized Ridge model 2a assigns non-zero weights to all highlighted variables, reflecting their continuously decreasing nature, while the Lasso model 2b promotes greater parsimony by acting by nullifying coefficients; finally, the Elastic Net model 2c has the characteristic of combining both strategies, retaining relevant variables even in the presence of multicollinearity.

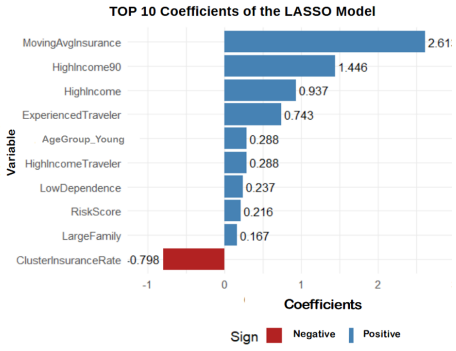
The coefficients estimated by the regularized models demonstrated good performance in identifying relevant variables. We highlight the variable ChronicByAge, which was demonstrated to be important in two of the three models. In the Ridge model, this variable presented the highest coefficient, considering absolute values, (-30.78), indicating a strong negative association with the basic variable. L2 regularization showed that the explanatory variable ChronicByAge is relevant. The value of this coefficient suggests that individuals with a greater accumulation of age-related comorbidities are less likely to purchase insurance.

The results observed for the Lasso model show that ChronicByAge does not appear among the top 10 predictors. This indicates that the L1 penalty imposed a stricter fil-

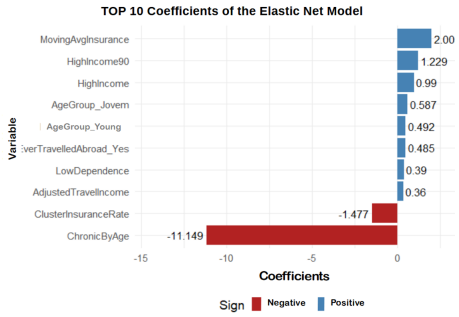
ter when compared to Ridge, highlighting variables with more isolated effects and less collinearity. In the Elastic Net model, which combines L1 and L2 regularizations, the variable chronic diseases weighted by age reappears with a substantial effect (-11.15). It can then be inferred that its absence in the Lasso model may be due to strong multicollinearity with other variables in the model. In addition, variables such as MovingAgInsurance, HighIncome00, ExperiencedTraveler and AgeGroup stand out among the main positive coefficients in the three models, showing a directly proportional influence. These variables are associated with a greater probability of purchasing travel insurance, reinforcing the role of income, previous travel experience and age as influencers in the profile of travelers inclined to purchase insurance.



(a) Ridge



(b) Lasso

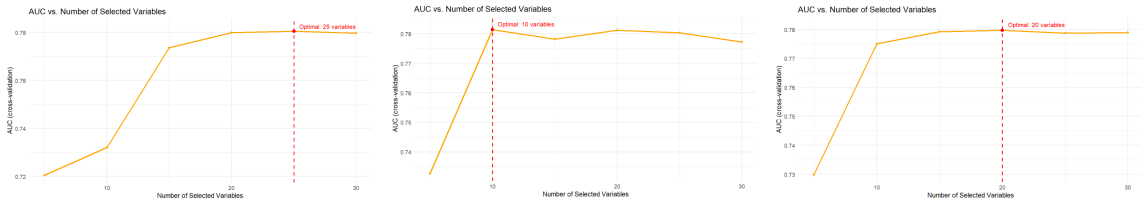


(c) Elastic Net

Figure 2: Top 10 most relevant variables based on the magnitude of the estimated coefficients in each regularized model based on AUC increment.

Next, figure 3 shows the evolution of the AUC value as a function of the number of selected variables, based on the importance criteria derived from the regularized models. Evaluating the Ridge model 3a, an improvement in performance is observed until reaching stabilization with a larger set of variables compared to the others. The Lasso model 3b, reaches a high AUC value in a few cycles considering the number of variables, while 3c presents an intermediate curve, due to its balance between the L1 and L2 penalties, as more variables are included.

Figure 3: AUC performance curves as a function of the number of selected variables for the Ridge, Lasso, and Elastic Net models. The results were obtained using stratified cross-validation.



(a) Ridge (b) Lasso (c) Elastic Net

In order to evaluate the metrics of the regularized models, the table 1 presents the confusion matrices of the three regularized models when using a training/testing split (80:20), highlighting the correct and incorrect classifications in the binary prediction task of the test set.

All models showed similar sensitivity, demonstrating equivalent capacity in correctly identifying the positive cases tested. The Lasso model demonstrated to have the highest specificity and precision among those evaluated, having a more accurate classification factor for this data set, as seen in the table 2 with an accuracy of 0.757. Ridge showed lower precision, despite a similar AUC, while the Elastic Net maintained intermediate performance as a hybrid solution between parsimony and robustness.

Table 1: Confusion matrices for the Ridge, Lasso, and Elastic Net models with True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) numbers.

	Ridge		**Lasso**		**Elastic Net**	
	Positive (S)	**Negative (N)**	**Positive (S)**	**Negative (N)**	**Positive (S)**	**Negative (N)**
Predicted Positive (S)	187 (TP)	97 (FN)	187 (TP)	97 (FN)	187 (TP)	97 (FN)
Predicted Negative (N)	39 (FP)	216 (TN)	34 (FP)	221 (TN)	37 (FP)	218 (TN)

Table 2 summarizes the overall performance metrics. The results corroborate previous observations: the Ridge model achieved the highest *recall* and AUC, the Lasso

model had the best *precision*, and the Elastic Net achieved intermediate performance. These findings demonstrate the impact of the choice of the regulation model on several indicators relevant to classification models using transformation.

Table 2: Evaluation metrics for the Ridge, Lasso, and Elastic Net models on the test set.

Model	AUC	Accuracy	Sensitivity (Recall)	Specificity	Precision (PPV)	F1-score	Balanced Accuracy
Ridge	0.809	0.7477	0.6585	0.8471	0.8274	0.7330	0.7528
Lasso	0.807	0.7570	0.6585	0.8667	0.8462	0.7402	0.7626
Elastic Net	0.809	0.7514	0.6585	0.8549	0.8348	0.7362	0.7567

3.2 Black-Box Models

In this section we will present a comparison where all the steps dimensionality reduction, selection of best features and hyperparameter optimization methods were performed using only black box models, that is, without considering regularization techniques here in order to have a comparative means between joint and isolated methods.

Table 3: **Performance and Optimized Hyperparameters of Black-Box Models in Travel Insurance Prediction**

Algorithm	Vars	AUC	Precision	Recall	F1 Score	Hiperparameters
Random Forest	12	0.8874	0.8022	0.8431	0.8222	ntree = 185, mtry = 10
XGBoost	8	0.8978	0.7788	0.9529	0.8571	max_depth = 3, eta = 0.0526, nrounds = 239
H2O GBM	10	0.8985	0.8135	0.8039	0.8087	max_depth = 3, ntrees = 310
LightGBM	13	0.9023	0.8097	0.8510	0.8298	learning_rate = 0.0246, nrounds = 256, num_leaves = 64
CatBoost	11	0.8986	0.7799	0.9451	0.8546	learning_rate = 0.0240, depth = 6, iterations = 403

Table 3 presents the results of the black-box models in the test set. The LightGBM model obtained the best performance compared to the others (AUC = 0.902), also with good indices in precision (0.81) and recall (0.85), resulting in an F1 Score (0.83). Next, we have CatBoost, which also appeared competitive (AUC = 0.899, F1 Score = 0.85 and sensitivity = 0.95). The XGBoost model had the second highest F1 Score (0.86) with a high recall (0.95), but a low precision (0.78) compared to the others, which suggests a tendency

towards a higher incidence of false positives. Random Forest presented consistent recall (0.84), but had lower AUC (0.887) and precision (0.80), slightly compromising its F1 Score (0.82). When evaluating the H2O GBM model, a lower performance is observed compared to the others (AUC = 0.899, F1 Score = 0.81 and recall = 0.80), but still with good metrics. These results suggest that models such as LightGBM and CatBoost have better adherence to the data set compared to the others.

4 Combined Approach

This section presents the comparative results between regularized models (Ridge, LASSO and Elastic Net) and black-box models (Random Forest, XGBoost, H2O GBM, LightGBM and CatBoost) applied to the task of predicting whether or not to purchase travel insurance. The objective is to analyze how different modeling strategies, with different levels of interpretability, affect the identification of relevant variables and the predictive performance of the final model. In the regularized models, variable selection was determined by L1 and/or L2 penalties, while in the black-box models, the importance of variables was determined based on internal metrics, such as information gain and noise reduction that did not generate performance gains. For both approaches, the most relevant variable subsets were collected, and hyperparameter tuning was performed through random search with 5-fold stratified cross-validation, each subset was reconditioned in the pipeline aiming at maximizing the AUC. Table 4 summarizes the main results of this analysis.

Table 4: Performance of Black-Box Models with Feature Selection via Regularization

Algorithm	Regularization	AUC	Precision	Recall	F1-Score	Hiperparameters
CatBoost	ElasticNet	0.8523	0.7412	**0.8635**	0.7975	learning_rate = 0.2043, depth = 4, iterations = 63
CatBoost	Lasso	**0.8611**	**0.8024**	0.8141	**0.8082**	learning_rate = 0.0242, depth = 4, iterations = 389
CatBoost	Ridge	0.8598	0.7568	0.8177	0.7859	learning_rate = 0.1666, depth = 5, iterations = 124
H2O GBM	ElasticNet	0.8609	0.7653	0.8100	0.7871	num_leaves = 4, learning_rate = 139
H2O GBM	Lasso	0.8565	0.7693	0.7439	0.7562	num_leaves = 4, learning_rate = 167
H2O GBM	Ridge	0.8565	0.7633	0.7889	0.7757	num_leaves = 4, learning_rate = 139
LightGBM	ElasticNet	0.8598	0.7552	0.8135	0.7832	learning_rate = 0.0196, max_depth = 337, eta = 52
LightGBM	Lasso	0.8588	0.7623	0.8058	0.7834	learning_rate = 0.1130, max_depth = 63, eta = 64
LightGBM	Ridge	0.8543	0.7523	0.7096	0.7301	learning_rate = 0.0780, max_depth = 78, eta = 52
Random Forest	ElasticNet	0.8484	0.7353	0.8415	0.7831	iterations = 141, depth = 3
Random Forest	Lasso	0.8503	0.7327	**0.8442**	0.7842	iterations = 283, depth = 6
Random Forest	Ridge	0.8513	0.7380	0.8337	0.7833	iterations = 393, depth = 7
XGBoost	ElasticNet	0.8545	0.7503	0.7771	0.7634	max_depth = 394, num_leaves = 3, learning_rate = 0.0565
XGBoost	Lasso	0.8583	0.7531	0.7917	0.7719	max_depth = 214, num_leaves = 5, learning_rate = 0.0721
XGBoost	Ridge	0.8584	0.7464	0.8535	0.7961	max_depth = 399, num_leaves = 3, learning_rate = 0.0317

Table 4 compares the performance of black-box models combined with different variable selection strategies via regularization (Elastic Net, LASSO, and Ridge). The combination of CatBoost with Lasso obtained the highest AUC (0.861) and the highest F1 Score (0.80) compared to the other models, with good indices for precision (0.8024) and recall (0.8141). This association suggests in our findings that the power of omitting irreversible features of the regularized LASSO model and the predictive power of the CatBoost classifier model are comparable.

Other hybrid modeling methods were also evaluated, including H2O GBM with Elastic Net, which obtained an AUC of 0.8609 and F1 Score of 0.7871, highlighting the highest recall (0.8183) among all the combination modeling methods listed. Following the table analysis, the XGBoost model combined with Ridge demonstrated a robust overall performance with an AUC of 0.8584 and an F1 score of 0.7961, demonstrating the effectiveness of Ridge even in highly nonlinear scenarios.

Meanwhile, LightGBM and Random Forest performed more moderately. LightGBM demonstrated consistency across regularizations, especially with Lasso (AUC = 0.8588, F1 = 0.7834), while Random Forest excelled mainly in its recovery (Elastic Net: 0.8415;

Lasso: 0.8442), but presented comparatively lower AUC scores.

Table 5: Performance and Hyperparameters – CatBoost with Lasso Regularization

Metric	**Value**
Algorithm	CatBoost
Regularization	LASSO
AUC	**0.8611**
Precision	**0.8024**
Recall	**0.8141**
F1-Score	**0.8082**
Hiperparameters	
learning_rate	0.0242
depth	4
iterations	389

Table 5 presents the performance of the CatBoost model with variable selection via Lasso regularization. This model had the highest AUC (0.861) among all combinations tested here, indicating relevant class discrimination. Of note are the hyperparameters obtained via optimization (learning rate = 0.0242, depth = 4, and 389 booster iterations). These results reinforce the quality of the association between CatBoost's learning capabilities and Lasso's feature selection.

5 Simulation Studies

5.1 Modeling and Experimental Evaluation

In this study, we evaluate the performance of 23 associated regression models, including three regularized linear models and five nonlinear algorithms, trained on the complete set of variables. We also present in this group of 23, fifteen hybrid models that combine variable selection with nonlinear algorithms again applied to the complete dataset. The models were trained using the full set of covariates generated from the

function proposed by [28], expressed as:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ and the independent variables $x_j \sim \mathcal{U}(0,1)$. Each penalized model produces a subset of variables, which is then used as input to five black-box algorithms: Random Forest, XGBoost, LightGBM, CatBoost, and H2O GBM. Each algorithm is trained three times, once for each subset of variables selected by Ridge, Lasso, and Elastic Net, resulting in 15 distinct hybrid models. Additionally, the same five algorithms are trained using all original variables without prior selection, resulting in a total of 23 distinct models. The models were trained using stratified cross-validation to ensure robustness of the root mean square error (RMSE) metrics.

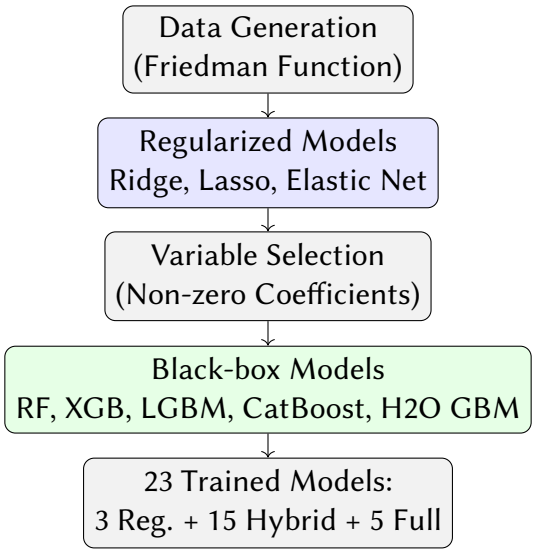


Figure 4: Pipeline for building hybrid models with variable selection via regularization.

5.2 Simulation Results

To evaluate the robustness and generalizability of the proposed models, we conducted a simulation study with different sample sizes ($n = 200$, 500, and 1000) and total numbers of predictor variables ($p = 5$, 10, 50). When the number of potential predictors was greater than 5, i.e., above that provided in the model being inferred, the extra predictors were generated independently from uniform distributions and did not affect the data generation mechanism, being merely noise predictors. The goal of this was to understand whether the algorithms were able to discriminate predictors that were truly important to predict the response. Each scenario included regularized models, black-box models using all variables, and hybrid models combining variable selection through penalized regression with machine learning algorithms.

Figure 5 presents the results for $n = 200$. It is evident that hybrid models based on CatBoost, LightGBM, and H2O GBM, combined with Elastic Net or Lasso, outperform, on average, both models trained on all variables and regularized linear models. The performance difference is particularly noticeable in low-dimensional scenarios ($p = 5$), suggesting that variable selection is especially beneficial when fewer informative predictors are available.

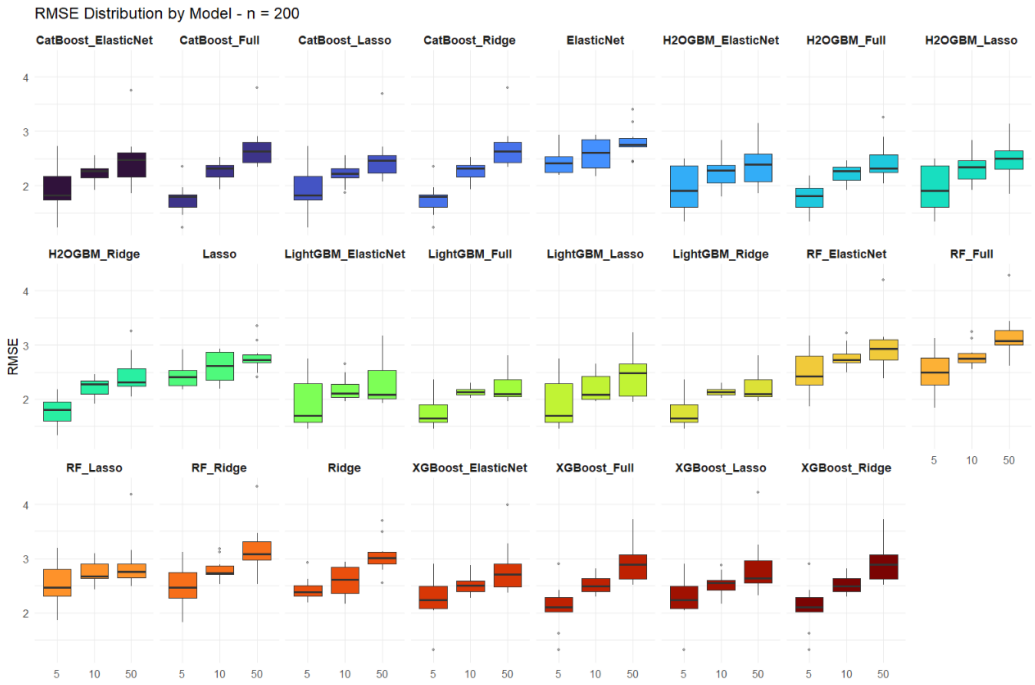


Figure 5: RMSE distribution by model – $n = 200$ **(a)**

Figure 6, corresponding to the intermediate case ($n = 500$), confirms the advantage of hybrid models, with *CatBoost_ElasticNet* and *H2OGBM_Lasso* achieving consistently low RMSE, even as dimensionality increases. Purely regularized models (*Ridge*, *Lasso*, *Elastic Net*) begin to lose competitiveness as p increases, exposing their limitations in capturing complex nonlinear interactions.

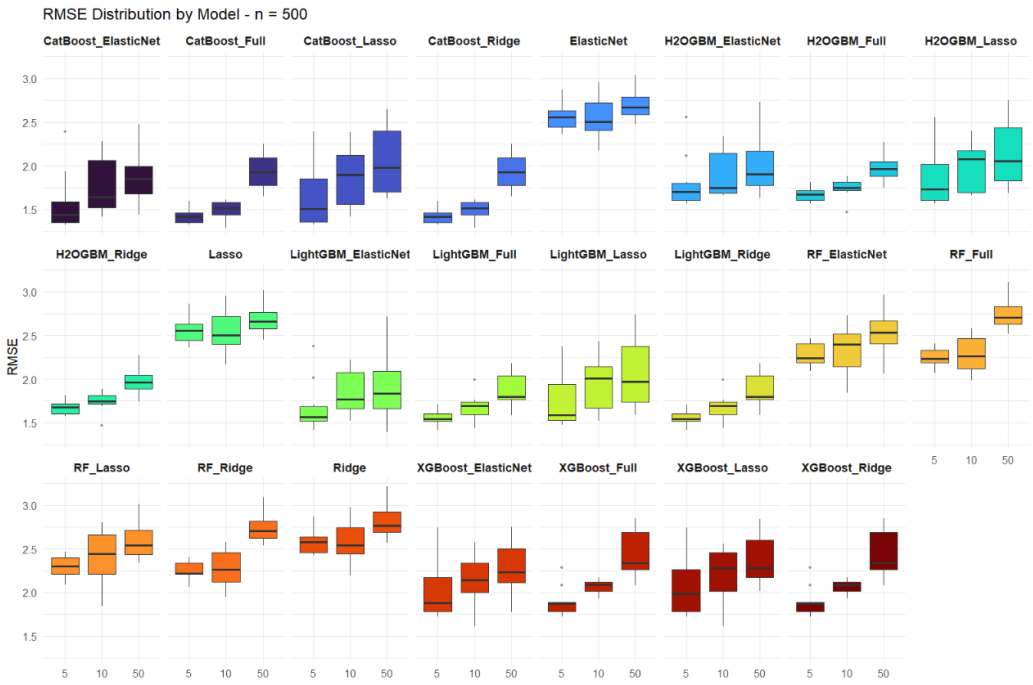


Figure 6: RMSE distribution by model – $n = 500$ **(b)**

Figure 7 shows the results for $n = 1000$, the previous observation remains the same since the hybrid models achieve the lowest mean errors at almost all levels of complexity (variations of p). As the sample size increases, the black-box estimators stabilize, but the hybrid models still stand out for their superior balance between bias and variance, particularly in controlling the variability of the RMSE.

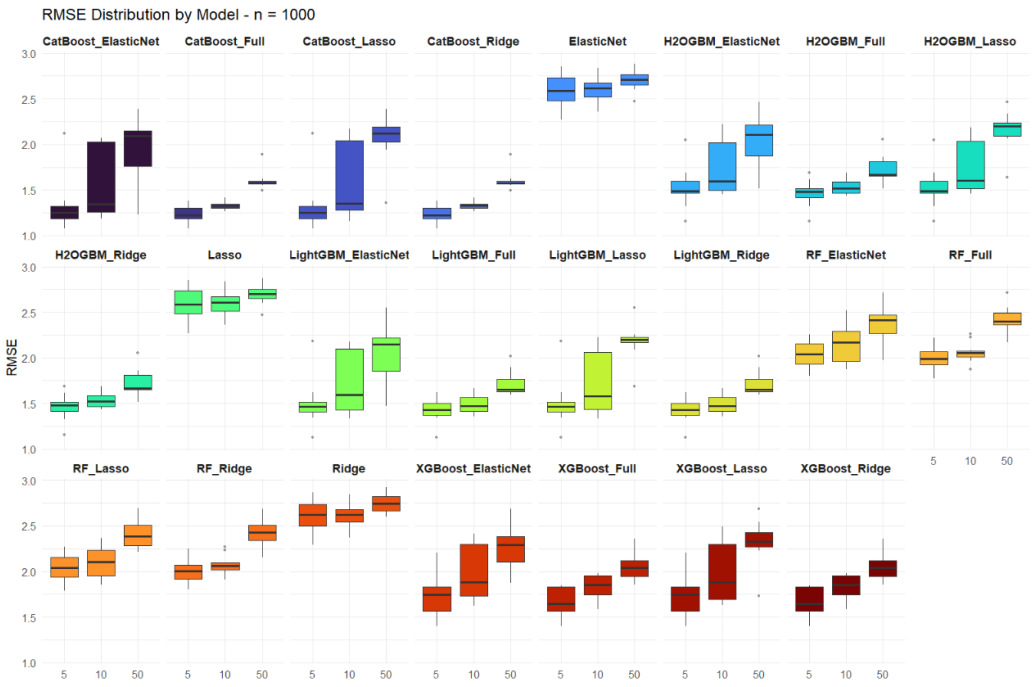


Figure 7: RMSE distribution by model – $n = 1000$ **(c)**

The 8 summarizes the average RMSEs among the simulated models. The black-box models without consortium present a good predictive power but do not differ substantially in the image from their hybrid equivalents (Regularized + Black-Box Equivalent).

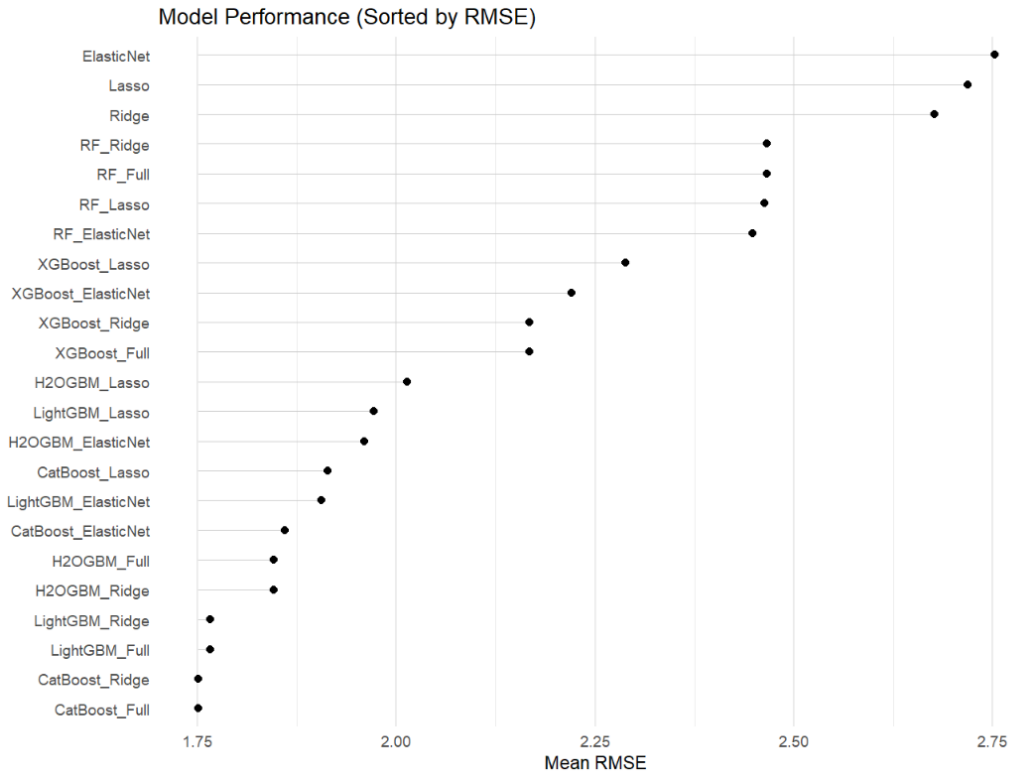


Figure 8: Comparison of predictive models ranked by their mean Root Mean Squared Error (RMSE). Lower values indicate better performance.

Simulation results indicate that combining penalized variable selection with Black-box algorithms is an effective strategy for improving predictive performance. Hybrid models strike a balance between robustness and accuracy, making them particularly suitable for applications such as predicting travel insurance adoption, where the number of important and irrelevant variables can vary significantly.

6 Conclusion

Our results demonstrate that the performance of the methodologies is distinct from each other, with gains in specific aspects, whether precision or interpretability, and important empirical nuances of the sector were observed, depending on the focus, among

the strategies considered.

The regularized linear models presented AUC values around 0.810 and F1 scores ranging from 0.73 to 0.74, indicating reliable predictive power, but with relatively inferior performance. It is worth noting that these models offer greater interpretability, which continues to be a critical advantage in contexts involving regulatory compliance, as is the case in the insurance sector. This is an environment where having extensive documentation and validation by experts is essential to launch any product in the insurance market.

In comparison, the pure black-box models, trained with the complete set of variables, obtained superior performance, in terms of F1 (which balances precision and recall). The LightGBM model (AUC = 0.902, F1 = 0.82) is recognized in the literature and by data bureaus in the insurance market for its great predictive power. As widely discussed, this model (and the others of this nonlinear profile) is suitable for production systems where real-time decision making and low error tolerance are essential. The hybrid models showed competitive performance, but slightly lower than their full-variable counterparts. The best among them was CatBoost with Lasso, achieving AUC = 0.8611 and F1 = 0.8082. Although this combination, in this application case and more generally in our simulations, slightly reduces the predictive power in the AUC, it offers better interpretability, reduced dimensionality and computationally faster inference times, important considerations in high-throughput environments, constrained service load times or when model transparency is required. In this scenario, hybrid models are a viable application option for the travel insurance sector and new work is encouraged in order to expand the possible scope of action of this methodology, which can be used in regression and classification problems.

References

[1] C. N. das Seguradoras (CNseg), "Procura por seguro viagem dispara no primeiro bimestre de 2024," https://cnseg.org.br/noticias/procura-por-seguro-viagem-dispara-no-primeiro-bimestre-de-2024, 2024, acesso em: maio 2025.

[2] Accenture, "Tendências na indústria de seguros 2023," https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-Insurance-Trends-2023-Brazil-V2.pdf, 2023, acesso em: maio 2025.

[3] S. T. Lim, J. Y. Yuan, K. W. Khaw, and X. Chew, "Predicting travel insurance purchases in an insurance firm through machine learning methods after covid-19," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 43–58, 2023.

[4] X. Li, "Exploring the potential of machine learning techniques for predicting travel insurance claims: A comparative analysis of four models," *Academic Journal of Computing & Information Science*, vol. 6, no. 4, pp. 118–125, 2023.

[5] R. Sahai, A. Al-Ataby, S. Assi, M. Jayabalan, P. Liatsis, C. K. Loy, A. Al-Hamid, S. Al-Sudani, M. Alamran, and H. Kolivand, "Insurance risk prediction using machine learning," *Lecture Notes in Networks and Systems*, vol. 165, pp. 419–433, 2023.

[6] Z. Sun, "Research on changes in travel insurance premiums driven by climate change: A case study of hong kong region," *Journal of Environmental Management*, vol. 325, pp. 116–123, 2024.

[7] X. Tian, J. Todorovic, and Z. Todorovic, "A machine-learning-based business analytical system for insurance customer relationship management and cross-selling," *Journal of Applied Business and Economics*, vol. 25, no. 6, pp. 256–270, 2023.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, and W. Ma, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] J. Friedman, T. Hastie, and R. Tibshirani, "glmnet: Lasso and elastic-net regularized generalized linear models," 2024, r package version 4.1-5. [Online]. Available: https://cran.r-project.org/web/packages/glmnet/index.html

[12] R. C. Team, "R: A language and environment for statistical computing," 2024, r Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[13] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[16] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

[17] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[18] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[19] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.

[20] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1–22, 2010.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013.

[24] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in neural information processing systems*, vol. 31, 2018, pp. 6638–6648.

[26] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, vol. 30, 2017, pp. 4765–4774.

[28] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

7 Anexo I

Table 6: Descrição das variáveis utilizadas no modelo preditivo

Variável (em português)	**Descrição**	**Categoria**	**Origem**
Idade	Idade do cliente	Numérica contínua	Original
Renda anual	Renda bruta anual declarada	Numérica contínua	Original
Número de familiares	Total de membros da família	Numérica discreta	Original
Graduação	Indica se é graduado	Categórica binária	Original
Tipo de emprego	Vínculo empregatício declarado	Categórica nominal	Original
Doenças crônicas	Número de condições crônicas declaradas	Numérica discreta	Original
Participação em milhagem	Participa de programa de milhas	Categórica binária	Original
Já viajou ao exterior	Histórico de viagem internacional	Categórica binária	Original
Contratou seguro	Variável-alvo (0 = não, 1 = sim)	Categórica binária	Original
Renda per capita	Renda anual dividida pelo número de familiares	Numérica contínua	Derivada
Alta renda	Acima do percentil 75 de renda	Categórica binária	Derivada
Idade normalizada	Idade padronizada (z-score)	Numérica contínua	Derivada
Doença crônica elevada	Acima da mediana de doenças crônicas	Categórica binária	Derivada
Frequência de viagem	Indicador baseado em milhagem e viagem internacional	Categórica binária	Derivada

Variável (em português)	Descrição	Categoria	Origem
Emprego privado	Indicador binário de setor privado/autônomo	Categórica binária	Derivada
Baixa dependência familiar	Até 3 membros na família	Categórica binária	Derivada
Renda por idade	Renda dividida pela idade do cliente	Numérica contínua	Derivada
Faixa etária	Categorização por grupos etários	Categórica ordinal	Derivada
Viajante com alta renda	Alta renda com histórico de viagens	Categórica binária	Derivada
Renda top 10%	Acima do percentil 90 de renda	Categórica binária	Derivada
Renda per capita normalizada	Padronização da renda per capita	Numérica contínua	Derivada
Viajante experiente	Já viajou internacionalmente	Categórica binária	Derivada
Família numerosa	Acima do 3º quartil em membros familiares	Categórica binária	Derivada
Doenças por idade	Razão entre doenças crônicas e idade	Numérica contínua	Derivada
Score de propensão	Escore baseado em renda, doenças e viagem	Numérica contínua	Derivada
Dependência financeira	log(Família / Renda)	Numérica contínua	Derivada
Score de viagem	Escore baseado em fidelidade e viagens	Numérica discreta	Derivada
Tempo de experiência	Idade estimada de início da vida profissional	Numérica contínua	Derivada
Emprego estável	Indicador para setor público	Categórica binária	Derivada
Renda ajustada à viagem	Renda ajustada para viajantes frequentes	Numérica contínua	Derivada
Score de risco	Composto de doenças, idade e viagem	Numérica contínua	Derivada
Score de risco normalizado	Padronização do score de risco	Numérica contínua	Derivada
Cluster de perfil	Agrupamento com base em idade, renda e emprego	Categórica ordinal	Derivada
Média por cluster	Probabilidade média de seguro no cluster	Numérica contínua	Derivada
Média móvel por grupo	Contratação média por grupo etário	Numérica contínua	Derivada