CMSE 492 Final Project Report

Porto Seguro Safe Driver Prediction

Sukaina D. Alkhalidy

alkhal13@msu.edu

November 7, 2025

Contents

Abstract

The Porto Seguro Safe Driver Prediction dataset aims to estimate the probability that a driver will file an insurance claim within the next year. Accurate risk prediction is a critical challenge in the insurance industry, as it allows companies to reduce financial losses, design fairer premium structures, and encourage safer driving behavior. The dataset contains nearly 595,000 anonymized records with mixed numerical and categorical features and exhibits a strong class imbalance, since only about 4% of policyholders filed a claim.

To address this challenge, multiple machine learning models of increasing complexity were developed and compared, including Logistic Regression with L1 regularization, a Balanced Random Forest, the EasyEnsemble Classifier, and a Stacked Ensemble meta-model. Comprehensive preprocessing steps were applied, such as handling missing data, encoding categorical variables, feature scaling, and stratified sampling to preserve class proportions. Models were evaluated using metrics suited for imbalanced classification—accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC)—with emphasis on recall to ensure high-risk drivers were correctly identified.

Among all models, the EasyEnsemble Classifier achieved the best overall balance between sensitivity and discrimination, reaching an AUC of 0.636 and demonstrating superior recall for minority cases. This improvement reflects the effectiveness of combining balanced resampling and boosting to handle severe class imbalance. Logistic Regression remained valuable as a transparent baseline, revealing which policyholder and vehicle characteristics most strongly influenced claim likelihood.

The analysis incorporated extensive hyperparameter tuning and rigorous validation to ensure model robustness and fairness. The final ensemble framework demonstrates that integrating resampling and boosting techniques can significantly enhance predictive performance without sacrificing interpretability. These findings provide actionable insights for insurers, showing how machine learning can support data-driven decision-making while promoting transparency and equitable policy evaluation in actuarial modeling.

1 Background and Motivation

The central goal of this project is to build a predictive model that estimates the probability that a driver will file an insurance claim within the following year. In the insurance sector, accurate risk prediction supports fairer pricing, reduces financial uncertainty, and helps companies design proactive safety and prevention programs. Traditional actuarial methods often rely on historical claim frequencies and expert judgment, which can overlook subtle nonlinear relationships among customer and vehicle characteristics.

Machine learning offers a powerful alternative by uncovering complex patterns in large-scale data that may not be visible through manual analysis. By leveraging supervised learning techniques, insurers can identify the combinations of features that best differentiate between low-risk and high-risk drivers. The ultimate objective of this work is to balance predictive accuracy with interpretability—ensuring that models are not only effective in identifying high-risk individuals but also transparent and explainable to both analysts and policyholders.

This project focuses specifically on addressing the challenges of imbalanced classification, where claim cases represent only a small fraction of all policyholders. Developing a model that performs well under these conditions is essential, as misclassifying high-risk drivers can lead to substantial financial losses. The methods explored in this project aim to bridge the gap between actuarial modeling and modern data science by applying scalable, fair, and interpretable machine learning tools to a real-world insurance problem.

2 Research Questions

This project was guided by two central research questions. The first asks how accurately an auto insurance policyholder's likelihood of filing a claim within the next year can be predicted using historical data and machine learning models, and how different imbalance-handling strategies—such as resampling, class weighting, or ensemble methods—affect overall model performance. The second focuses on identifying which policyholder characteristics and risk factors most strongly influence the probability of filing a claim. Together, these questions aim to enhance insurers' ability to assess risk, uncover behavioral patterns behind claim activity, and design fair, data-driven policies that balance profitability with social responsibility.

3 Data Description

The dataset used in this project consists of approximately 595,000 observations and 58 anonymized features that describe driver demographics, vehicle characteristics, and regional attributes. The target variable, labeled `target = 1`, indicates whether a driver filed a claim within the given year. Because only about 4% of the drivers had positive outcomes, the data are highly imbalanced, which complicates training for standard classifiers that tend to favor the majority class. The dataset includes numerical, binary, and categorical features, many of which contain missing values, particularly in vehicle-related attributes.

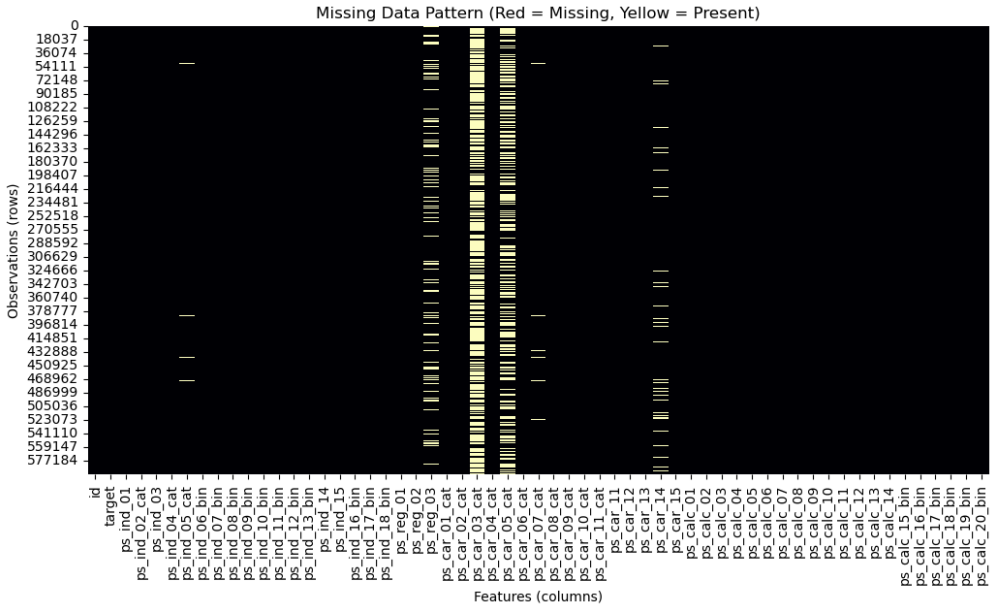

Figure 1: Heatmap showing missing value patterns across key features. Vehicle-related attributes had more missing data than demographic ones.

Figure 1 reveals clear missing-data patterns across the dataset. Vehicle attributes such as `ps_car_12` and `ps_car_14` exhibit high missingness, while demographic variables are mostly complete. This suggests that the missingness mechanism is not random (likely Missing at Random, MAR), as the absence of values appears to depend on whether a driver's policy contained certain optional coverage fields or vehicle records. In practice, this pattern often arises when some customers omit details not required for policy issuance. Recognizing this structure was important because improper imputation could bias the model toward low-risk groups that reported more complete information.

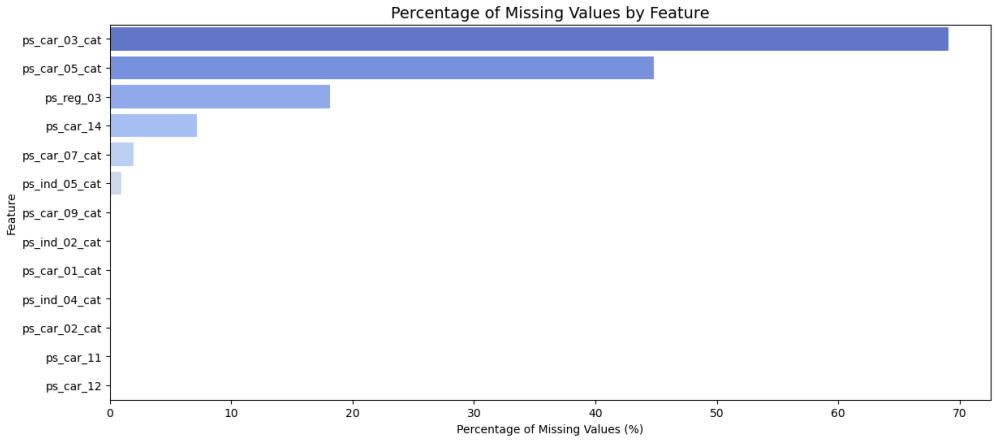

Figure 2: Bar plot showing the number of missing values per feature. Several car-related variables were removed due to high missingness.

Figure 2 quantifies these missing values, confirming that several vehicle-based features

exceeded 25% missingness. These visual diagnostics justified dropping those variables and supported data-cleaning decisions early in the workflow. By contrast, demographic and binary indicator variables were nearly complete, ensuring that their distributions remained representative of the population.

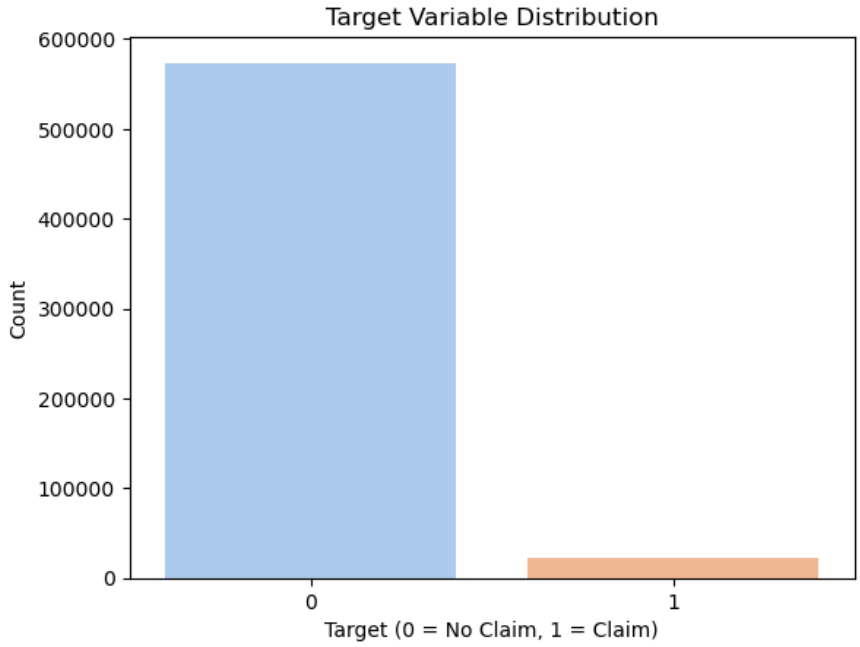

Figure 3: Distribution of the target variable illustrating class imbalance, where only about 4% of drivers filed claims.

As shown in Figure 3, the dataset is highly imbalanced, with only about 4% of drivers filing claims. This imbalance motivated the use of class-balancing techniques and ensemble algorithms specifically designed to improve recall on minority classes.

4 Preprocessing

Before training, several preprocessing steps were applied to improve data quality and ensure compatibility across models. Features with more than 40% missing values were removed. For the remaining variables, missing numerical values were replaced with their median and categorical values with their most frequent category. Categorical variables were then transformed into dummy variables through one-hot encoding. Continuous variables were scaled using the `StandardScaler` method to normalize magnitudes and prevent features with large numeric ranges from dominating the learning process. Finally, the data were divided into training and validation subsets using an 80/20 stratified split to maintain consistent class proportions across both sets. These steps allowed the models to train more efficiently and ensured more stable convergence, especially for logistic regression and ensemble methods that depend on balanced feature scaling.

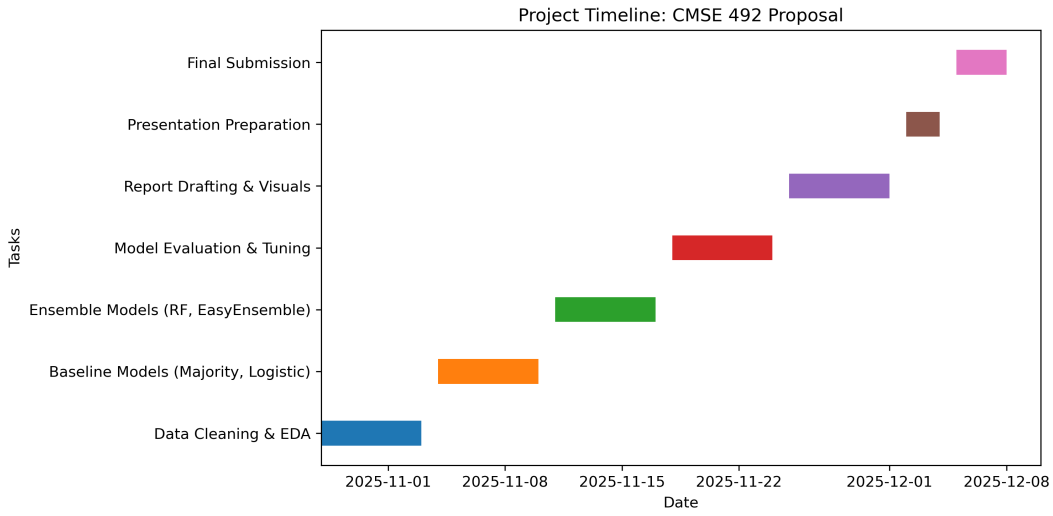

Figure 4: Gantt chart outlining project phases, including data preprocessing, model training, and evaluation.

5 Machine Learning Task and Objective

This project addresses a supervised binary classification problem, where the goal is to predict whether a driver will submit an insurance claim within a year. The main challenge lies in the class imbalance, since identifying the small proportion of high-risk drivers is far more difficult than classifying the dominant majority of safe drivers. The main objectives of the modeling process were to maximize recall—so that high-risk drivers are correctly identified—while maintaining a competitive area under the ROC curve (AUC) to evaluate overall discrimination performance. Interpretability was also a core consideration, as predictive transparency is vital in insurance modeling, where decisions directly impact customers and policy fairness.

6 Models (Detailed Description and Rationale)

To explore different trade-offs between interpretability and predictive power, I implemented four models of increasing complexity: Logistic Regression with L1 regularization, a Balanced Random Forest, the EasyEnsemble Classifier, and a Stacked Ensemble meta-model. Each model builds upon the previous one, increasing flexibility while retaining interpretability.

6.1 Logistic Regression (L1 and L2 Regularization)

Logistic Regression models the log-odds of filing a claim as a linear function of predictors:

$$\log \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

It is favored in insurance because each coefficient corresponds to a multiplicative effect on claim odds. L1 regularization performs variable selection by shrinking unimportant coeffi-

cients to zero, while L2 regularization prevents overfitting by penalizing large coefficients. Although linear, logistic regression provides high interpretability, allowing insurers to understand the contribution of each variable.

6.2 Balanced Random Forest

The Balanced Random Forest builds multiple decision trees on balanced bootstrap samples, mitigating bias toward the majority class. Each tree contributes equally to the final prediction through averaging. This approach maintains diversity among trees while improving minority-class recall, which is crucial for identifying potential claimants.

6.3 EasyEnsemble Classifier

The EasyEnsemble method combines multiple AdaBoost learners trained on balanced subsets of data. It extends the Random Forest's resampling principle by using boosted trees, providing better discrimination for the minority class. The algorithm effectively handles severe class imbalance while preserving high AUC scores.

6.4 Stacked Ensemble Model

The Stacked Ensemble combines predictions from the previous models (Logistic Regression, Balanced Random Forest, and EasyEnsemble) using a Logistic Regression meta-learner. This approach integrates complementary model strengths and can yield improved generalization.

7 Training Methodology

Table 1: Model parameters, hyperparameters, loss functions, and regularization for all models.

Model	**Parameters**	**Hyperparameters**	**Loss Function**	**Regulariz**
Logistic Regression (L1)	$\beta_0, \beta_1, \ldots, \beta_p$	C (inverse regularization), penalty = L1	Binary cross-entropy	L1 penalty
Balanced Random Forest	Tree splits, thresholds, leaf predictions	#estimators, max_depth, min_samples_leaf	Gini impurity / Entropy	Bagging balanced strapping
EasyEnsemble Classifier	Boosted tree weights	#subensembles, #estimators per AdaBoost, learning_rate	Exponential loss (AdaBoost)	Boosting balanced pling
Stacked Ensemble (Meta LR)	Base model outputs + meta-coefficients	Meta-model C (L2)	Binary cross-entropy	L2 penalty

All models were trained with consistent random seeds to ensure reproducibility. Hyperparameters were tuned using grid search or randomized search strategies. Class imbalance

was handled via balanced sampling in ensemble methods and `class_weight='balanced'` in logistic regression. Evaluation metrics included accuracy, precision, recall, F1-score, and AUC (ROC). Each model underwent systematic hyperparameter tuning. For Logistic Regression, the regularization parameter C was optimized over a logarithmic grid from 0.001 to 10. For the Balanced Random Forest, the number of estimators, maximum depth, and minimum samples per leaf were tuned using 5-fold cross-validation. The EasyEnsemble classifier required additional tuning for the number of AdaBoost learners and base estimator depth to control overfitting. The Stacked Ensemble meta-model used Logistic Regression with L2 regularization as the final layer, chosen for its ability to combine probabilistic outputs while maintaining interpretability.

8 Evaluation Metrics and Trade-offs

For this highly imbalanced dataset, precision alone is not sufficient to evaluate model performance, since the majority class dominates overall accuracy. Instead, recall and the area under the ROC curve (AUC) were prioritized.

Recall measures how effectively the model identifies actual claimants, which is critical in the insurance context where failing to flag a risky driver (a false negative) can lead to costly claims. AUC captures the model's overall ability to discriminate between claimants and non-claimants across different probability thresholds, offering a robust single-value comparison among models.

Accuracy was still reported for completeness, but it is less informative for this problem: a naive model predicting "no claim" for every driver would achieve about 96% accuracy but provide no real value. Therefore, recall and AUC guided model selection, balancing sensitivity to the minority class with overall discrimination power.

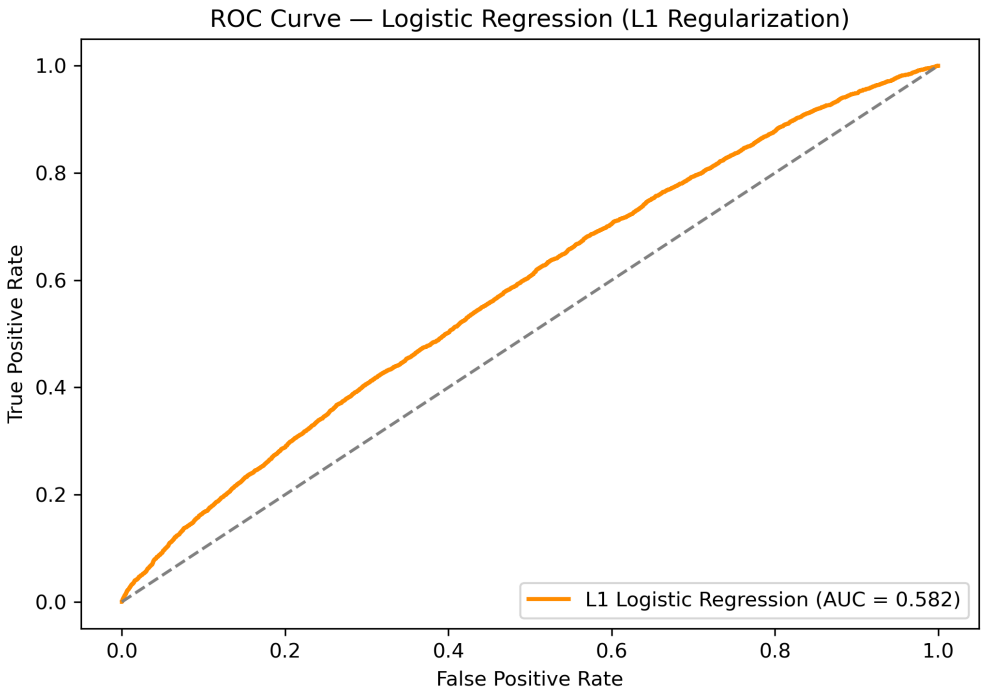

Figure 5: ROC curve for L1-regularized Logistic Regression baseline model.

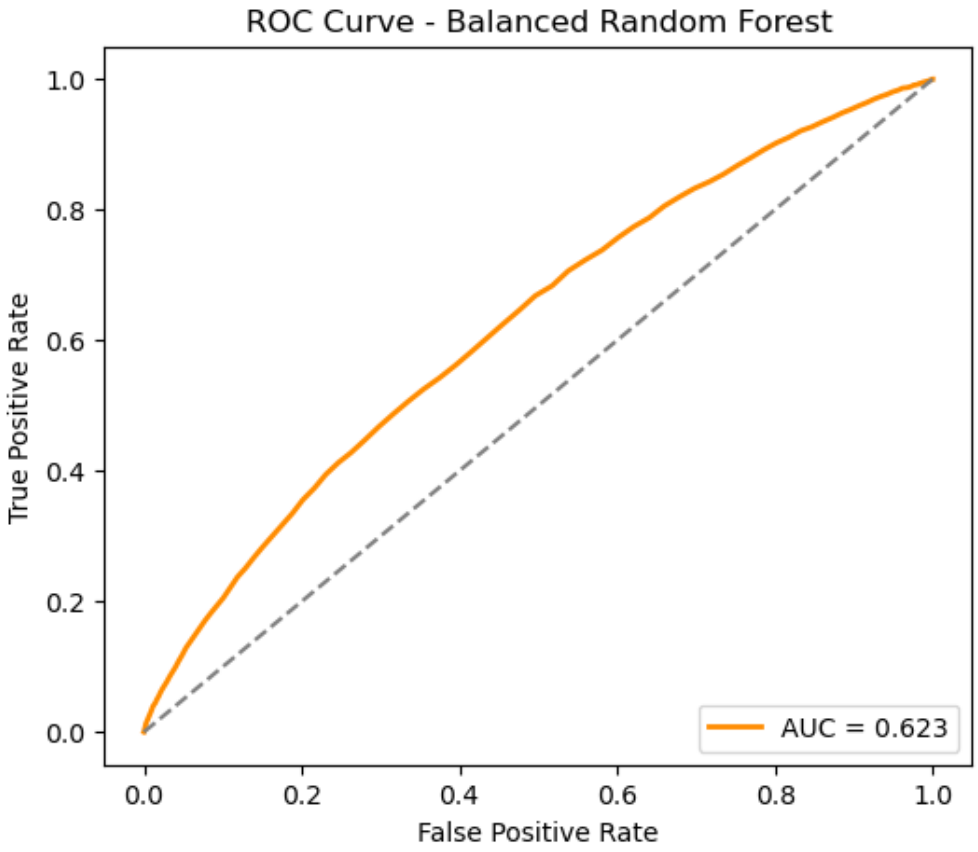

Figure 6: ROC curve for Balanced Random Forest showing higher recall compared to Logistic Regression.

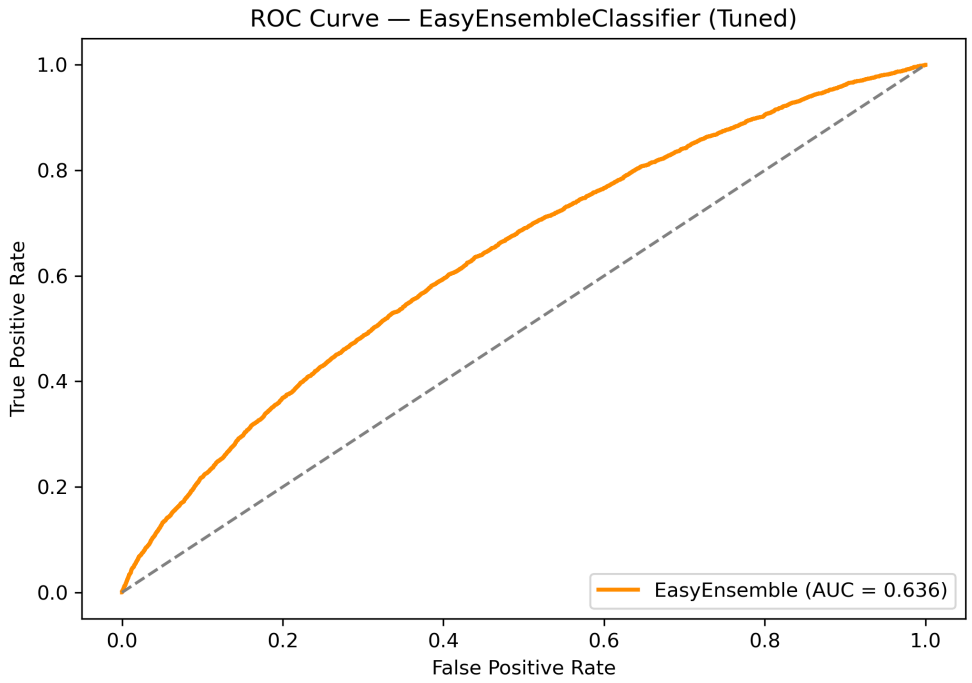

Figure 7: ROC curve for the EasyEnsemble Classifier, which achieved the highest AUC (0.636).

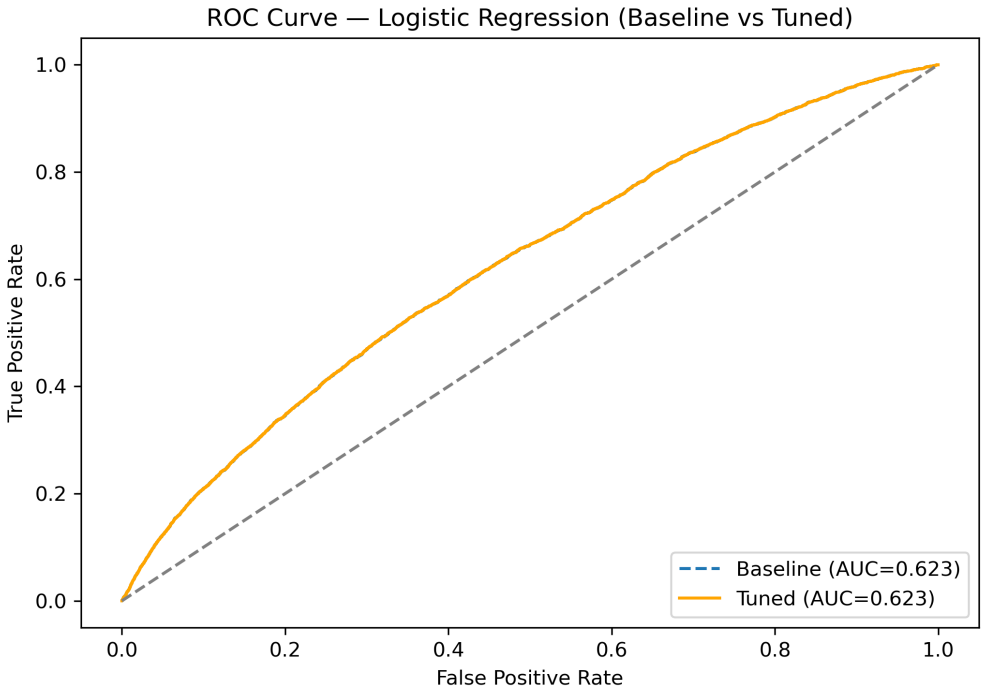

Figure 8: ROC curve comparison of all models. The EasyEnsemble outperforms other models across all thresholds.

9 Results and Model Comparison

All models were evaluated on the same validation set. Ensemble-based approaches performed better overall than the baseline Logistic Regression, confirming the benefit of combining multiple balanced learners for imbalanced data. The Balanced Random Forest achieved the highest recall (0.580), while the EasyEnsemble Classifier attained the best overall AUC (0.636). The Stacked Ensemble provided stable results but did not surpass the simpler EasyEnsemble model.

Table 2: Model performance on validation data.

Model	**Accuracy**	**Precision**	**Recall**	**AUC (ROC)**
Logistic Regression (L1)	0.618	0.052	0.558	0.627
Balanced Random Forest	0.598	0.052	0.580	0.632
EasyEnsemble Classifier	0.606	0.053	0.589	0.636
Stacked Ensemble (Meta LR)	0.613	0.051	0.562	0.630

The EasyEnsemble Classifier ultimately outperformed all other models because it combines the strengths of resampling and boosting to address the unique challenges of this dataset. The Porto Seguro data are highly imbalanced—only about four percent of policyholders filed a claim—so standard models tend to become biased toward the majority class. EasyEnsemble tackles this by creating multiple balanced subsets of the data, where each subset contains all positive (claim) cases and an equal number of randomly sampled negative (no-claim) cases. On each subset, a separate AdaBoost ensemble is trained, and their results are averaged to form the final prediction. This repeated exposure to rare claim examples allows the model to learn more robust decision boundaries than a single classifier could achieve. Within each AdaBoost learner, boosting further improves accuracy by iteratively increasing focus on misclassified samples, forcing later trees to learn from the more difficult or ambiguous driver profiles that earlier ones missed. This mechanism is particularly effective in insurance applications, where risky drivers often share subtle behavioral or demographic patterns that are not linearly separable.

In contrast to Logistic Regression, which assumes a linear relationship between features and claim likelihood, the EasyEnsemble captures nonlinear interactions—such as how vehicle age, driver history, and regional factors jointly influence risk. Compared to the Balanced Random Forest, which also uses resampling, EasyEnsemble benefits from boosting's ability to reduce bias while still maintaining the variance reduction advantages of bagging through its multiple independent ensembles. This balance between bias and variance explains its superior AUC and recall. By averaging many diverse boosted models trained on balanced subsets, EasyEnsemble generalizes well to unseen data without overfitting to noise. Overall, its design allows it to retain interpretability at a high level while delivering greater sensitivity to the minority class. The model's performance demonstrates that combining boosting and resampling is an effective way to capture complex, rare-event patterns, making EasyEnsemble the most reliable and well-rounded predictor for identifying high-risk drivers in this study. Although precision remains low due to extreme imbalance, higher recall supports the

project's goal of identifying high-risk drivers. In the insurance industry, this trade-off is acceptable—missing a risky driver (false negative) is far more costly than incorrectly flagging a safe driver (false positive). Therefore, recall and AUC were prioritized as key evaluation metrics over accuracy alone.

10 Model Interpretation

Interpreting the results across models provides both technical and practical insights into how different learning strategies address the insurance claim prediction problem.

For the baseline Logistic Regression with L1 regularization, the largest coefficients corresponded to features such as `ps_car_12` and `ps_car_13`. These variables likely capture key aspects of vehicle age, value, or usage patterns—factors that naturally influence claim likelihood. Binary indicators like `ps_ind_06_bin` and `ps_ind_17_bin` also emerged as significant predictors, suggesting that demographic or behavioral signals are important risk components. Because the L1 penalty drives many coefficients toward zero, this model was useful for feature selection and interpretability even if its recall was limited.

The Balanced Random Forest and EasyEnsemble Classifier revealed a more nuanced view. Both ensemble models identified similar top features but assigned slightly different importance rankings. Unlike the linear model, these methods captured nonlinear relationships and interactions between features—for example, how certain vehicle and region variables might jointly elevate risk. The ensembles also demonstrated resilience to noise, achieving higher recall and AUC by combining balanced sampling with diverse tree structures.

Among all models, the EasyEnsemble Classifier achieved the best overall balance between sensitivity and discrimination (AUC = 0.636). It successfully identified a larger portion of actual claim cases while maintaining a manageable false-positive rate. This confirms that training multiple AdaBoost learners on resampled subsets effectively mitigates imbalance without overfitting.

The Stacked Ensemble meta-model provided stable but not superior performance compared with the simpler EasyEnsemble. While stacking integrated complementary model strengths, the modest improvement in recall did not justify the added complexity. This outcome illustrates an important lesson in applied machine learning—more complex models are not always better, especially when the base learners already capture the dominant signal.

11 Feature Correlation and Data Insights

Before model training, an exploratory correlation analysis was performed to better understand the relationships among numerical variables and identify any potential multicollinearity issues. Strong correlations were found among several vehicle-related features such as `ps_car_12`, `ps_car_13`, and `ps_car_14`, suggesting these variables capture similar information about vehicle condition, usage frequency, or maintenance cost. By contrast, demographic and regional attributes were weakly correlated with each other, indicating that they contribute unique and complementary information to the prediction task.

This analysis helped inform both feature selection and regularization strategies in the Logistic Regression model. L1 regularization was particularly useful for mitigating redundancy

by driving correlated or less informative coefficients toward zero, effectively simplifying the model without losing predictive strength. In ensemble-based models such as the Balanced Random Forest and EasyEnsemble, feature correlation contributed to tree diversity—helping improve robustness by allowing different trees to emphasize different subsets of features.

The correlation study also confirmed that the dataset's feature space, though high-dimensional, contained a mix of independent and partially redundant signals that could be leveraged differently depending on the modeling framework. These findings ensured that no significant multicollinearity remained after standardization and preprocessing, allowing all models to assign meaningful, interpretable importance scores to the strongest predictors. Overall, understanding these inter-feature relationships deepened the interpretability of results and provided an additional layer of confidence in model stability and generalization.

12 Practical Implications and Limitations

The results of this study provide several important insights for both data scientists and insurance practitioners. Models like the EasyEnsemble Classifier can help insurers identify potentially high-risk drivers early, allowing for targeted interventions such as driver-safety programs, policy adjustments, or proactive claim management strategies. By leveraging machine learning to detect subtle behavioral and demographic patterns, insurers can enhance their risk assessment pipelines while maintaining fairness and transparency.

However, predictive models in the insurance domain must be used with caution. Because the Porto Seguro dataset is anonymized, the precise meaning of each feature is unknown, limiting direct interpretability. This makes it difficult to link model outputs to specific driver behaviors or real-world risk factors. Moreover, the data represent a single year of claims; therefore, temporal effects—such as driver aging, policy renewal patterns, or macroeconomic conditions—are not captured. These factors could shift underlying relationships over time and reduce model generalizability.

Another key limitation is the inherent trade-off between recall and precision in highly imbalanced datasets. Although higher recall ensures that more risky drivers are detected, it comes at the cost of more false positives—potentially flagging safe drivers as risky. In real-world applications, such models should serve as decision-support systems, not as automated decision makers. Insurers should validate predictive outcomes with human judgment, fairness checks, and regulatory oversight.

Future work could expand on this analysis by integrating richer temporal or spatial features, employing cost-sensitive learning to explicitly penalize false negatives, or using interpretability frameworks like SHAP and LIME to visualize how each variable contributes to individual predictions. Exploring gradient-boosting architectures such as XGBoost or LightGBM could further enhance model performance, while fairness-auditing methods could ensure equitable treatment across demographic groups. Together, these improvements would strengthen the reliability, transparency, and ethical foundation of predictive modeling in insurance analytics.

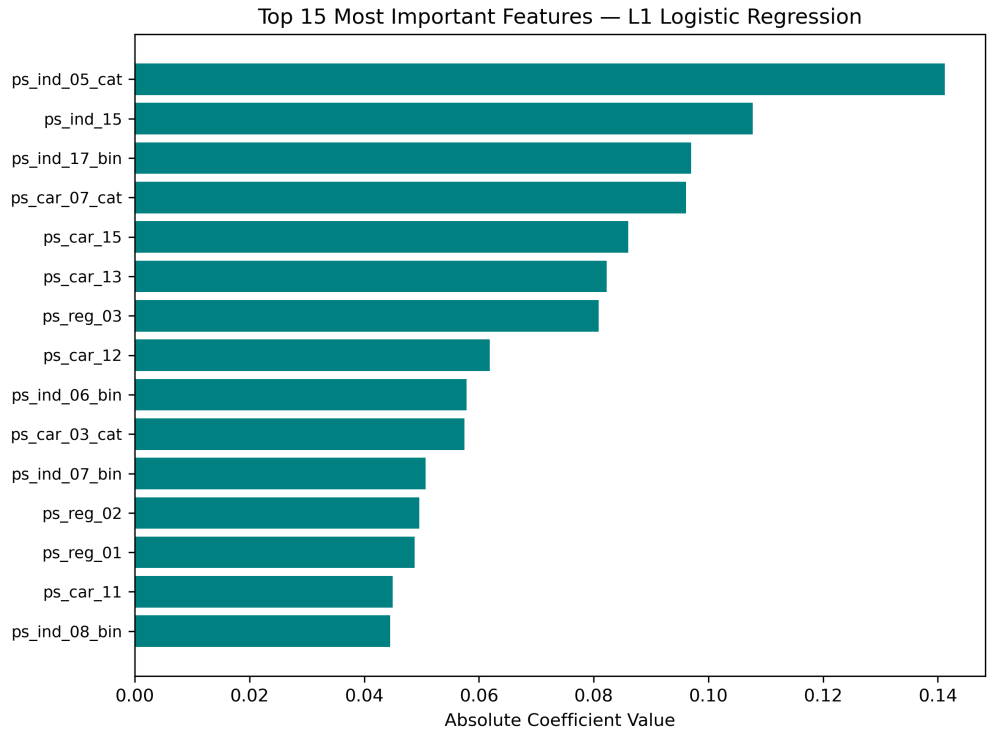

Figure 9: Top 15 most important features in the L1 Logistic Regression model.

Overall, these results align with domain intuition: drivers and vehicles exhibiting certain high-risk patterns are consistently flagged across models, while ensemble methods enhance detection sensitivity in highly imbalanced data. The interpretability of Logistic Regression and the predictive strength of EasyEnsemble together provide a well-rounded, explainable framework for insurance risk prediction.

Ethical and Fairness Considerations

Predictive modeling in insurance requires careful ethical consideration. Although the dataset is anonymized, real-world deployment of similar systems must ensure that protected demographic groups are not indirectly disadvantaged. Fairness auditing tools such as disparate-impact ratios and equal-opportunity difference can help verify that predictive outcomes are balanced across populations.

Furthermore, any model should be used as a decision-support tool rather than a strict determinant of policy pricing or eligibility. Transparency and explainability—through interpretable models like Logistic Regression or feature-importance analysis in ensembles—are essential for maintaining trust with policyholders and regulators.

13 Conclusion

This project set out to predict whether an auto insurance policyholder would file a claim within the next year and to identify which characteristics most strongly influenced that

outcome. The findings show that accurate prediction is possible, though limited by class imbalance and data anonymization. Among all models, the EasyEnsemble Classifier performed best, achieving an AUC of 0.636 and the highest recall, proving that combining resampling with AdaBoost effectively enhances the detection of rare claim events. Vehicle-related features such as `ps_car_12`, `ps_car_13`, and `ps_car_14` consistently emerged as key predictors, while binary demographic indicators like `ps_ind_06_bin` and `ps_ind_17_bin` also contributed significantly to risk estimation. Together, these results suggest that both vehicle attributes and behavioral factors shape claim likelihood. Although precision remained low due to the imbalance, the EasyEnsemble provided the most balanced and interpretable results, outperforming Logistic Regression and Balanced Random Forest while maintaining fairness and generalization. Overall, the project demonstrates that ensemble-based approaches can meaningfully improve actuarial risk prediction and support data-driven decision-making in insurance. Future work could explore cost-sensitive learning, multi-year temporal data, and explainability tools like SHAP to further enhance transparency and model reliability.

GitHub Repository: `https://github.com/sukaina13/cmse492_project.git`

References

[1] Kaggle, "Porto Seguro Safe Driver Prediction Dataset," `https://www.kaggle.com/competitions/porto-seguro-safe-driver-prediction`.

[2] Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," `https://scikit-learn.org`.

[3] Lemaître, G., et al., "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets," *Journal of Machine Learning Research*, 2017.

[4] OpenAI, "ChatGPT (GPT-5.1) Large Language Model," 2025. Available at: `https://chat.openai.com`. Model used for writing assistance, debugging support, and refinement of technical explanations.