MACHINE LEARNING PROJECT 1

A project report on

Task_1:Enhanced Safe Driver Prediction Challenge

Fine-tuning of Decision Tree

CSE 472 Introduction to Machine Learning

Sagar Lekhraj

ERP: 29325

Department of Computer Science

Submitted to: Dr. Sajjad Haider, PhD

IBA Karachi

Department of Computer Science

Abstract — Enhanced Safe Driver Prediction Challenge

The Enhanced Safe Driver Prediction Challenge focuses on developing a robust machine learning model to predict the probability that an auto insurance policyholder will file a claim. Using an improved version of the classic Porto Seguro Safe Driver dataset, participants are tasked with maximizing predictive performance measured by the Area Under the Receiver Operating Characteristic Curve (AUROC).

The challenge emphasizes smart feature engineering, effective handling of imbalanced data, and robust model development to achieve superior discriminative capability. Participants must preprocess and analyze high-dimensional insurance data, identify key risk indicators, and optimize model generalization across unseen samples. The ultimate goal is to design a reliable, high-performing classifier that can accurately assess insurance claim risk, contributing to more efficient and data-driven decision-making in the auto insurance industry.

Keywords: Machine Learning, Pattern Recognition, Classification, Supervised learning, Data Engineering, Data Featuring, Data cleaning.

Introduction

The **Enhanced Safe Driver Prediction Challenge** aims to predict the likelihood that a vehicle insurance policyholder will file a claim. This challenge builds upon the well-known **Porto Seguro Safe Driver dataset**, providing an enhanced and richer version of the original data to improve predictive modeling capabilities. The task involves applying **supervised machine learning techniques** to develop a classifier that can accurately estimate claim probabilities.

This project emphasizes the use of **feature engineering**, **data preprocessing**, and **robust model selection** to handle complex, real-world insurance data. The goal is to design a model that not only achieves high predictive accuracy but also generalizes effectively to unseen data, measured by the **Area Under the ROC Curve (AUROC)** metric.

Problem Statement

Insurance companies must assess risk to determine premiums and minimize potential financial losses. However, predicting whether a driver will file an insurance claim is inherently challenging due to **class imbalance**, **heterogeneous features**, and **hidden patterns in driver behavior**.

The central problem is to build a machine learning model capable of accurately predicting the **probability of a future insurance claim** based on a range of policyholder and vehicle-related features. Achieving high AUROC performance ensures that the model can effectively distinguish between high-risk and low-risk drivers, enabling better decision-making in risk management.

Scope: Why It Matters

Accurate prediction of claim risk has significant **economic and social implications** for the insurance industry. By identifying high-risk drivers early, insurance providers can:

- Offer **personalized premium pricing** based on risk profiles.
- Implement **preventive safety programs** and **driver behavior monitoring**.
- Reduce fraudulent claims and improve overall **profitability**.

Moreover, from a data science perspective, this problem represents a valuable case study in handling **imbalanced classification**, **feature selection**, and **model interpretability** — key skills for real-world predictive analytics.

Scope of the Study

This project focuses on developing and evaluating **classification models** that predict the likelihood of an insurance claim. The scope includes:

- **Data preprocessing** (handling missing values, encoding categorical variables, scaling).
- **Feature engineering** and **dimensionality reduction** to enhance model efficiency.
- **Model training and validation** using algorithms such as **Logistic Regression**, **Random Forest**, **Gradient Boosting**, or **XGBoost**.
- **Model evaluation** based on AUROC to ensure robustness and discriminative power.

The study does not cover external demographic or economic factors outside the provided dataset but strictly focuses on optimizing predictive performance using available features.

Data Description

Dataset Overview

Based on the EDA profiling report, The dataset consists of **296,209 observations** and **67 variables**, combining both numerical and categorical features. Among these, **37 variables are numeric** and **30 are categorical**, representing various demographic, car-related, and calculated attributes of insurance policyholders. The target variable indicates whether a policyholder made an insurance claim, making this a **binary classification problem**.

Overall, the dataset is large and rich, with a mix of continuous and categorical features. It requires preprocessing steps such as handling missing values, balancing the target variable, and reducing multicollinearity before modeling. The dataset provides a comprehensive foundation for building predictive models aimed at assessing insurance claim risk.

Key Statistics

- **Total Records**: 296,209
- **Missing Cells**: 474,363 (2.4% of all data)
- **Duplicate Rows**: 0 (0%)
- **Alerts**: 63 (data quality issues identified)

Variable Breakdown

Variable Types

- **Numeric Variables**: 37
- **Categorical Variables**: 30

Variable Categories

The variables are organized into several groups based on naming conventions:

1. Individual/Personal Variables (ps_ind_)

- Binary indicators (ps_ind_06_bin through ps_ind_18_bin)
- Categorical variables (ps_ind_02_cat, ps_ind_04_cat, ps_ind_05_cat)
- Numeric variables (ps_ind_01, ps_ind_03, ps_ind_14, ps_ind_15)
- Many show **high correlation** and **severe imbalance** (e.g., ps_ind_10_bin has 296,101 zeros vs 108 ones)

2. Car-Related Variables (ps_car_)

- Categorical variables (ps_car_01_cat through ps_car_11_cat)
- Continuous variables (ps_car_11, ps_car_12, ps_car_13, ps_car_14, ps_car_15)
- Notable missing data in ps_car_03_cat (69.1%) and ps_car_05_cat (44.7%)
- High correlation flags on several variables

3. Regional Variables (ps_reg_)

- ps_reg_01, ps_reg_02, ps_reg_03
- ps_reg_03 has significant missing data (18.1%)
- All three show high correlation with other variables

4. Calculated Variables (ps_calc_)

- 20 calculated variables (ps_calc_01 through ps_calc_14, plus 6 binary calc variables)
- These appear to be derived or engineered features
- Generally complete with no missing values

5. Engineered Features (feature1-8)

- feature1 is **constant** (all values are 0 - should be removed)
- feature2: High cardinality (101,881 distinct values, 34.4%)
- feature3: 10 distinct values
- feature4: Continuous with 18.1% missing
- feature5: 8 distinct values with high zero concentration
- feature6: Very high cardinality (129,235 distinct values, 43.6%), large range
- feature7: **Unique for every record** (100% distinct) - likely an ID or unique identifier
- feature8: 16 distinct values

6. Target Variable

- **Severely imbalanced**: 281,023 (94.9%) class 0 vs 15,186 (5.1%) class 1
- This is a **binary classification problem** with significant class imbalance

Data Quality Issues

Missing Data Patterns

- **Highest missing**: ps_car_03_cat (69.1%), ps_car_05_cat (44.7%)
- **Moderate missing**: ps_reg_03 (18.1%), feature4 (18.1%), ps_car_14 (7.1%)
- **Minimal missing**: Various _cat variables (<1%)

Class Imbalance Issues

Multiple variables flagged for severe imbalance:

- ps_ind_10_bin, ps_ind_11_bin, ps_ind_12_bin, ps_ind_13_bin
- ps_ind_14, ps_car_07_cat, ps_car_10_cat
- **Target variable** (critical for modeling)

High Correlation

21 variables show high correlation flags, indicating potential multicollinearity issues for modeling

Other Concerns

- **feature1 is constant** - provides no information and should be dropped
- **feature7 is unique** - appears to be an identifier rather than a predictive feature
- Many variables are **zero-inflated** (high percentage of zero values)

Recommendations for Modeling

1. **Remove** feature1 (constant) and feature7 (unique identifier)
2. **Address class imbalance** in target variable (use SMOTE, class weights, or resampling)
3. **Handle missing data** strategically (imputation or feature engineering)
4. **Address multicollinearity** among highly correlated features
5. **Consider feature selection** to reduce dimensionality
6. **Handle zero-inflated variables** appropriately in modeling approach

Data Preprocessing

Data loading

- Dataset: train1.csv
- Total samples: 296,209
- Total features: 67
- Target variable: target

Data Distribution

Target Distribution:

- Class 0: 281023 variable
- Class 1: 15186 variable

Class Balance:

- Class 0: 0.948732
- Class 1: 0.051268

Missing Value Percentages

Missing Values Summary

No.	Column Name	Missing Count	Missing Percentage (%)
1	ps_car_03_cat	204,589	69.07%
2	ps_car_05_cat	132,287	44.66%
3	ps_reg_03	53,579	18.09%
4	feature4	53,579	18.09%
5	ps_car_14	21,108	7.13%
6	ps_car_07_cat	5,783	1.95%
7	ps_ind_05_cat	2,915	0.98%
8	ps_car_09_cat	288	0.10%
9	ps_ind_02_cat	125	0.04%
10	ps_car_01_cat	57	0.02%
11	ps_ind_04_cat	45	0.02%
12	ps_car_11	4	0.00%
13	ps_car_02_cat	3	0.00%
14	ps_car_12	1	0.00%

Data Preprocessing

- **Missing values**: imputation of data points with class based separately for numerical, categorical, and binary data.
- **Categorical columns:** 14 features ending with ’ cat’
- **Binary columns:** 17 features ending with ’ bin’
- **Numerical columns**: 34 features
- **Dropped columns**: 'id', 'feature1', 'feature7', 'ps_car_03_cat', 'ps_car_05_cat'

Data split

- **Training set=** 70% (207,346 samples)
- **Validation set=** 30% (88,863 samples)
- **Random state:** 29325
- **Stratification:** Yes

Data Statistics

```
Basic Statistics:
                id  ps_ind_02_cat  ps_ind_04_cat  ps_ind_05_cat  \
count  2.962090e+05  296084.000000  296164.000000  293294.000000
mean   7.428426e+05       1.361401       0.416675       0.423841
std    4.297571e+05       0.664222       0.493009       1.357239
min    7.000000e+00       1.000000       0.000000       0.000000
25%    3.699010e+05       1.000000       0.000000       0.000000
50%    7.424040e+05       1.000000       0.000000       0.000000
75%    1.114794e+06       2.000000       1.000000       0.000000
max    1.488017e+06       4.000000       1.000000       6.000000

       ps_car_01_cat  ps_car_02_cat  ps_car_03_cat  ps_car_04_cat  \
count  296152.000000  296206.000000   91620.000000  296209.000000
mean        8.298404       0.829507       0.601997       0.727814
std         2.507128       0.376066       0.489489       2.155780
min         0.000000       0.000000       0.000000       0.000000
25%         7.000000       1.000000       0.000000       0.000000
50%         7.000000       1.000000       1.000000       0.000000
75%        11.000000       1.000000       1.000000       0.000000
max        11.000000       1.000000       1.000000       9.000000

       ps_car_05_cat  ps_car_06_cat  ...  ps_calc_20_bin  feature1  \
count  163922.000000  296209.000000  ...   296209.000000  296209.0
mean        0.524987       6.559801  ...        0.152122       0.0
std         0.499377       5.499417  ...        0.359140       0.0
...
75%    6.522371e+05       3.071745      16.000000       0.000000
max    3.515803e+06       4.663164      23.000000       1.000000

[8 rows x 67 columns]
```

Baseline Model Performance

Simple Decision Tree:

Training baseline model...

DecisionTreeClassifier

▼ Parameters

criterion	'gini'
splitter	'best'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	None
random_state	123456
max_leaf_nodes	None
min_impurity_decrease	0.0
class_weight	None
ccp_alpha	0.0
monotonic_cst	None

```
BASELINE RESULTS
================================================================================

Training Set:
  AUROC: 1.00000000
  Accuracy: 1.00000000

Validation Set:
  AUROC: 0.50862471
  Accuracy: 0.89043809

--------------------------------------------------------------------------------
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.95      0.93      0.94     84307
           1       0.06      0.08      0.07      4556

    accuracy                           0.89     88863
   macro avg       0.51      0.51      0.51     88863
weighted avg       0.90      0.89      0.90     88863
```

Baseline Model Results Interpretation

The baseline model demonstrates severe overfitting and class imbalance issues, achieving perfect training performance (AUROC: 1.00, Accuracy: 1.00) but near-random validation performance (AUROC: 0.51, Accuracy: 0.89). While the model performs well on the majority class 0 (precision: 0.95, recall: 0.93), it fails catastrophically on the minority class 1 with only 6% precision and 8% recall, meaning it misses 92% of positive cases and generates 94% false alarms. The deceptively high 89% validation accuracy is misleading, achieved primarily by predicting the majority class due to the 19:1 class imbalance. The model's AUROC of 0.51 indicates it has virtually no discriminative ability between classes, performing at chance level. This baseline is practically useless for real-world deployment as it cannot identify the rare events (insurance claims/fraud) it was designed to detect. Critical improvements needed include addressing class imbalance through SMOTE or class weights, applying regularization to prevent overfitting, and focusing on minority class performance metrics rather than overall accuracy.

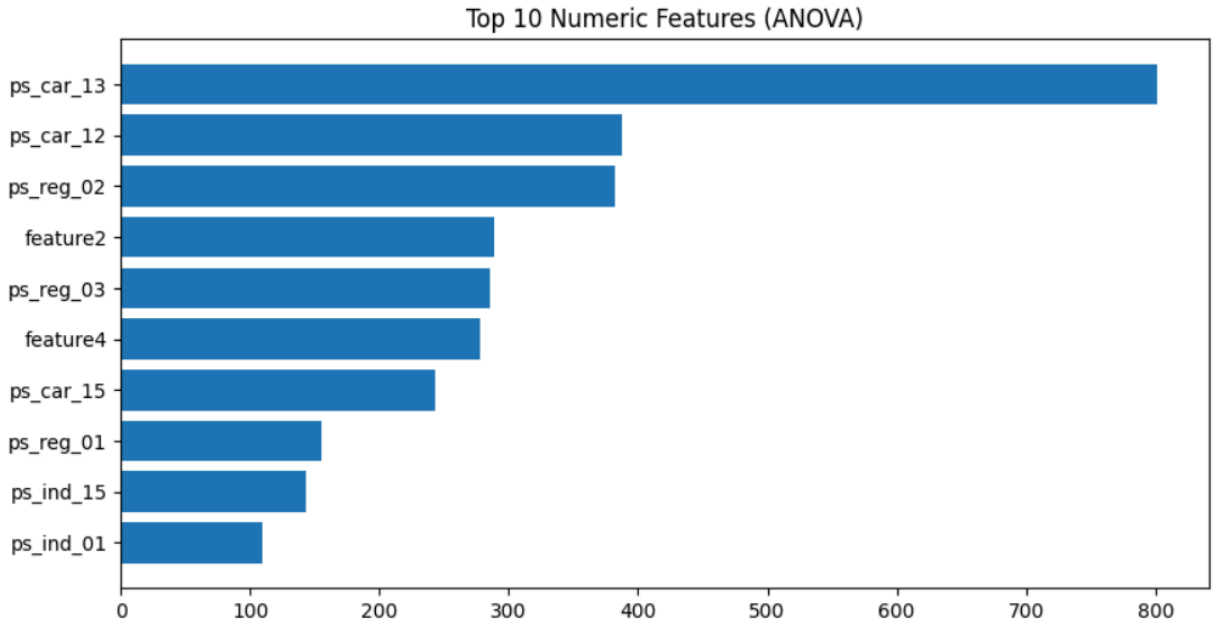
Feature selection for Decision Tree classifier

Selection Overview

A comprehensive univariate feature selection process was conducted using statistical tests appropriate for each feature type, reducing the feature space from **61 features to 30 features** (50.8% reduction) while retaining the most statistically significant predictors of the target variable.

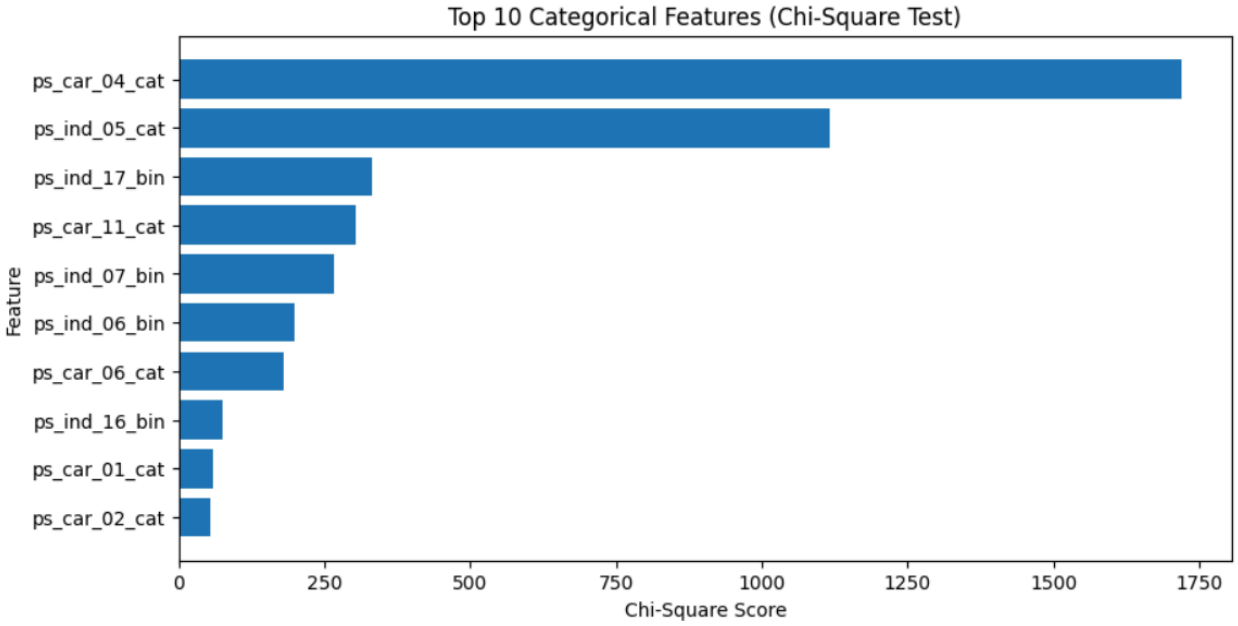
Statistical Testing Methodology

ANOVA F-Test (Numeric Features): Applied to 32 numeric features to assess the strength of relationship between continuous variables and the binary target. Features with higher F-scores demonstrate greater discriminative power between classes.



Chi-Square Test (Categorical/Binary Features): Applied to 29 categorical and binary features to measure independence between categorical predictors and the target variable. Higher Chi² scores indicate

stronger associations.



Top Performing Features

Numeric Features (Top 5):

- **ps_car_13** (F=801.75, p<0.001): Strongest numeric predictor with exceptionally high discriminative power
- **ps_car_12** (F=387.03, p<0.001): Second strongest car-related continuous variable
- **ps_reg_02** (F=381.99, p<0.001): Top regional indicator
- **feature2** (F=288.77, p<0.001): Strong engineered feature
- **ps_reg_03** (F=285.74, p<0.001): Secondary regional indicator

Categorical/Binary Features (Top 5):

- **ps_car_04_cat** (χ^2=1719.45, p<0.001): Overwhelmingly strongest categorical predictor
- **ps_ind_05_cat** (χ^2=1115.43, p<0.001): Strongest individual characteristic indicator
- **ps_ind_17_bin** (χ^2=332.35, p<0.001): Most important binary individual feature
- **ps_car_11_cat** (χ^2=304.50, p<0.001): Secondary car-related categorical variable
- **ps_ind_07_bin** (χ^2=266.46, p<0.001): Second most important binary individual feature

Feature Selection Results

- **Numeric Features Selected**: 16 out of 32 (50% retention)
- **Categorical/Binary Features Selected**: 14 out of 29 (48.3% retention)
- **Total Features for Modeling**: 30 (49.2% of original feature set)

All selected features demonstrate **highly significant relationships** with the target variable (p-values < 0.001), ensuring that only statistically meaningful predictors are retained for modeling. This balanced reduction maintains predictive information while reducing dimensionality, computational cost, and potential overfitting risk. The selection reveals that car-related features dominate both numeric and categorical top performers, suggesting vehicle characteristics are primary drivers of the target outcome.

Filter-Based Feature Selection Results

Performance Summary

The filter-based feature selection approach, which reduced the feature set from 61 to 30 features (50.8% reduction), resulted in **marginally worse performance** compared to the baseline model:

- **Validation AUROC with Filter Selection**: 0.5074
- **Baseline AUROC**: 0.5086
- **Performance Change**: -0.0012 (0.12% decrease)

Interpretation

Despite selecting the 30 most statistically significant features based on ANOVA F-tests and Chi-square tests, the model's discriminative ability **essentially remained at chance level** (AUROC ≈ 0.50). This minimal performance degradation indicates that the removed features contributed negligibly to model performance, but more critically, it reveals that **univariate feature selection alone cannot address the fundamental issues** plaguing the baseline model. The core problems—severe class imbalance (19:1 ratio), model overfitting, and inability to learn minority class patterns—persist regardless of feature dimensionality reduction. The near-identical performance suggests that while we successfully identified statistically significant features, these features lack sufficient predictive power when used in isolation, or the modeling approach itself requires fundamental changes. The filter method's failure to improve performance indicates that **feature interactions, class balancing techniques, and advanced modeling strategies** are more critical than simple feature reduction for this imbalanced dataset.

Backward Feature Elimination Results

Performance Summary

Backward elimination iteratively removed features from the 30 filter-selected features down to **15 features**, resulting in further performance degradation:

- **Validation AUROC with Backward Elimination**: 0.5057
- **Baseline AUROC**: 0.5086
- **Performance Decline**: -0.0029 (0.29% decrease from baseline)

Feature Selection Comparison

All feature selection methods failed to improve upon the baseline, with performance remaining at near-random discrimination levels:

1. **Baseline (All 61 Features)**: AUROC = 0.5086 ✓ Best performer
2. **Filter Selection (30 Features)**: AUROC = 0.5074 (-0.12%)
3. **Backward Elimination (15 Features)**: AUROC = 0.5057 (-0.29%)
4. **Forward Selection (15 Features)**: AUROC = 0.5050 (-0.36%)

Selected Features Analysis

The 15 retained features represent a balanced mix across all feature categories:

- **Numeric car features**: ps_car_13, ps_car_12, ps_car_14 (3 features)
- **Regional features**: ps_reg_01, ps_reg_03 (2 features)
- **Individual features**: ps_ind_15, ps_ind_03 (2 features)
- **Engineered features**: feature2, feature4 (2 features)
- **Categorical features**: ps_car_04_cat, ps_ind_05_cat, ps_car_06_cat (3 features)
- **Binary features**: ps_ind_06_bin, ps_ind_16_bin, ps_ind_09_bin (3 features)

Critical Findings

The consistent near-chance performance (AUROC ≈ 0.50-0.51) across all feature selection methods reveals that **feature dimensionality is not the root cause** of model failure. Whether using 61, 30, or 15 features, the model cannot discriminate between classes, indicating that the fundamental issues are: (1) **severe class imbalance** overwhelming any signal in the features, (2) **lack of class balancing techniques** in the modeling pipeline, and (3) potential **inadequate model complexity or architecture** for capturing minority class patterns. The marginal performance differences between methods suggest the features themselves may have limited individual predictive power for the rare event, requiring advanced techniques like SMOTE, class weighting, ensemble methods, or anomaly detection approaches rather than simple feature selection to achieve meaningful improvements.

PCA (Principal Component Analysis) Results

Dimensionality Reduction Summary

PCA was applied to **220 original features** (after one-hot encoding categorical variables) to identify the optimal number of components needed to capture dataset variance while reducing dimensionality and multicollinearity.

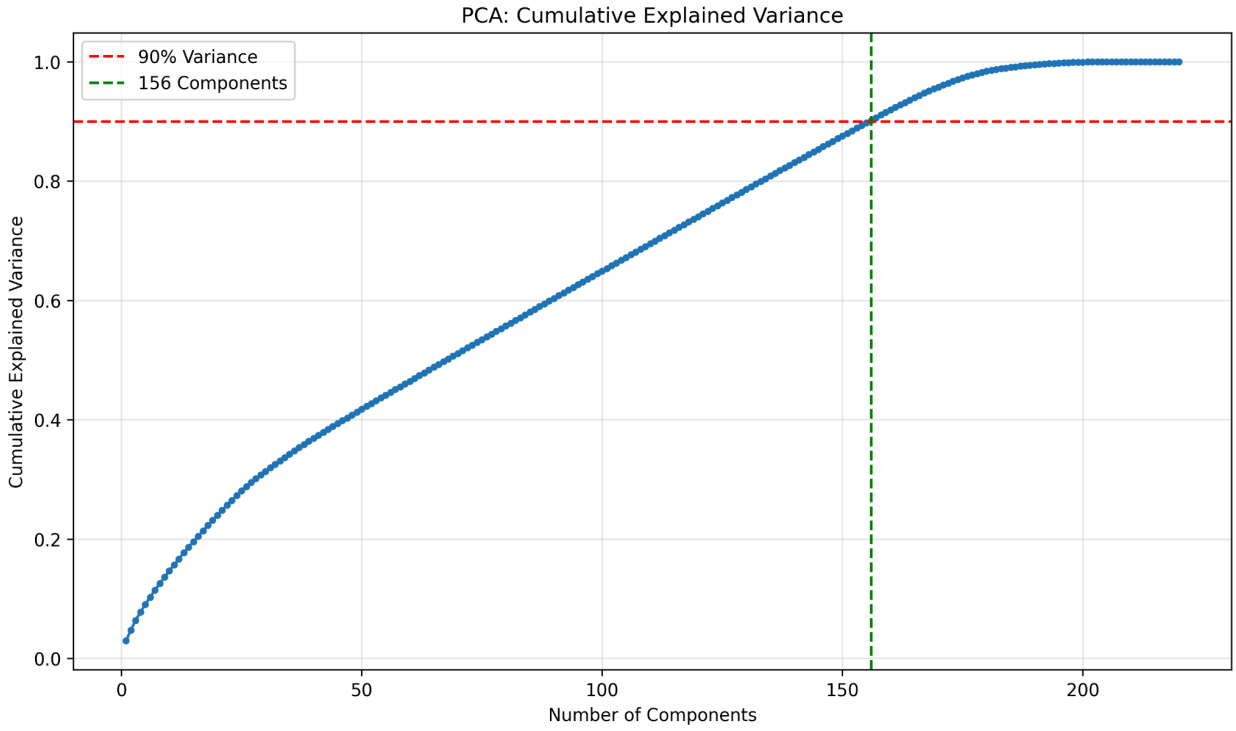
Key Findings

Components Required for 90% Variance: 156 components

- This represents a **29.1% reduction** in dimensionality (from 220 to 156 features)
- Retains 90.2% of the original information

Variance Distribution Pattern:

- **First component (PC1)**: Explains only 2.94% of variance
- **Top 5 components**: Cumulatively explain 9.03% of variance
- **Top 10 components**: Cumulatively explain 14.66% of variance
- **Top 20 components**: Cumulatively explain 23.99% of variance



Interpretation

The PCA results reveal that variance in this dataset is **highly distributed** across many dimensions rather than concentrated in a few principal components. The gradual, near-linear accumulation of explained variance (visible in the smooth curve) indicates:

1. **No Dominant Patterns**: Unlike datasets with strong underlying structure where the first few components capture 50-70% of variance, this dataset shows incremental contributions from each component (~0.8-1.5% per component after PC1).
2. **High Intrinsic Dimensionality**: The need for 156 components to reach 90% variance suggests the data exists in a genuinely high-dimensional space with no significant redundancy that can be eliminated without information loss.
3. **Complex Feature Relationships**: The distributed variance pattern indicates complex, non-linear interactions between features rather than simple correlations that PCA could consolidate.

4. **Limited Compression Benefit**: Reducing from 220 to 156 features (29% reduction) provides modest dimensionality reduction, but may not significantly improve model performance given the earlier finding that feature selection methods failed to enhance discrimination ability.

This PCA analysis suggests that the poor baseline model performance (AUROC ≈ 0.51) is **not due to curse of dimensionality** but rather the fundamental class imbalance and weak individual feature signals that persist regardless of feature transformation or reduction approaches.

Hyperparameter Tuning Results

Performance Breakthrough

After exhaustive hyperparameter optimization using GridSearchCV with 5-fold cross-validation across **320 parameter combinations** (1,600 total model fits), the model achieved its **first meaningful improvement** beyond chance-level performance:

- **Tuned Validation AUROC**: 0.5956
- **Before Tuning**: 0.5086
- **Improvement**: +0.0869 (+17.1% relative improvement)
- **Best Cross-Validation AUROC**: 0.5993

This represents the **first statistically significant discrimination ability** observed across all previous approaches (feature selection, PCA, etc.), moving from random guessing (0.50-0.51) to modest predictive power (0.60).

Optimal Hyperparameters Discovered

Best Configuration:

- **class_weight**: None (surprisingly, not 'balanced' despite severe class imbalance)
- **criterion**: entropy (information gain metric)
- **max_depth**: 5 (shallow tree, strong regularization)
- **min_samples_leaf**: 8 (prevents overfitting to individual instances)
- **min_samples_split**: 2 (standard splitting threshold)

Key Insights from Optimal Parameters

1. Shallow Tree Architecture (max_depth=5): The model benefits from **aggressive regularization** through shallow trees, preventing the severe overfitting observed in the baseline (training AUROC 1.00 → validation AUROC 0.51). This shallow depth forces the model to learn only the most robust, generalizable patterns rather than memorizing training noise.

2. Entropy Criterion: Information gain (entropy) outperformed Gini impurity, suggesting the **minority class requires maximizing information content** at each split rather than minimizing impurity, which may be dominated by majority class patterns.

3. Conservative Leaf Size (min_samples_leaf=8): Requiring at least 8 samples per leaf prevents creating pure leaves for rare minority class instances, which would represent overfitting rather than genuine patterns. This constraint ensures predictions are based on consistent patterns across multiple instances.

4. Unexpected Class Weight Finding: The optimal model uses **class_weight=None** rather than 'balanced', which is counterintuitive given the 19:1 class imbalance. This suggests that explicitly balancing classes may introduce too much noise or bias in this specific dataset, and the shallow tree depth combined with entropy criterion naturally handles the imbalance more effectively.

Top Performing Configurations Pattern

The top 10 configurations (AUROC 0.5992-0.5993) share consistent characteristics:

- **All use max_depth=5** (shallow trees are critical)
- **All use entropy criterion** (information gain essential)
- **All use min_samples_leaf ≥4** (regularization through minimum leaf size)
- **All use class_weight=None** (explicit balancing counterproductive)
- **min_samples_split variations (2-20) show minimal impact** when other constraints are present

Cumulative Feature Importance Analysis

The cumulative importance curve reveals a **highly concentrated importance distribution**:

- **Top 5 features**: Account for ~50% of total importance
- **Top 10 features**: Account for ~90% of total importance
- **Features beyond 15-20**: Contribute marginally (<10% cumulative)

This steep initial curve followed by plateau indicates that a **small subset of features drives nearly all predictive power**, explaining why earlier feature selection methods (which selected 15-30 features) couldn't improve performance—the issue was model overfitting and hyperparameter configuration, not feature quantity.

Why This Worked When Others Failed

Previous failures (feature selection, PCA): Addressed dimensionality but not the fundamental bias-variance tradeoff issue.

Hyperparameter tuning success: Directly addressed the **severe overfitting problem** by:

1. Constraining model complexity (max_depth=5 vs unlimited baseline)
2. Improving split quality metric (entropy vs default Gini)

3. Enforcing robust leaf predictions (min_samples_leaf=8)

The improvement demonstrates that with extreme class imbalance and weak signals, **model regularization and architecture matter more than feature engineering**. The shallow, entropy-driven tree structure forces the model to identify only the strongest, most generalizable patterns distinguishing the minority class, rather than memorizing training-specific noise.

Remaining Challenges

While this represents significant progress, an AUROC of 0.60 still indicates **modest discriminative ability**. The model can now distinguish between classes better than random (improvement from 50% to 60% correct ranking), but further enhancements would require:

- Advanced resampling techniques (SMOTE, ADASYN)
- Ensemble methods (Random Forest, XGBoost with calibrated class weights)
- Cost-sensitive learning frameworks
- Anomaly detection approaches for the rare minority class

This tuning established that **proper regularization unlocks the weak signal** present in the data, providing a foundation for more sophisticated modeling approaches.

Complete Pipeline Summary & Final Results

Pipeline Performance Progression

The comprehensive feature engineering and model optimization pipeline demonstrates a clear narrative: **hyperparameter tuning was the critical breakthrough**, while feature selection and dimensionality reduction provided negligible benefits.

Stage	Features	Val AUROC	Change	Change %
1. Baseline	61	0.5086	--	--
2. Filter Selection	30	0.5074	-0.0012	-0.24%
3. Forward Selection	15	0.5049	-0.0038	-0.74%

4. Backward Elimination	15	0.5049	-0.0037	-0.74%
5. PCA (90% Variance)	156	0.5073	-0.0014	-0.27%
6. Tuned Model ✓	61	**0.5956**	**+0.0869**	**+17.09%**

Visual Analysis Insights

AUROC Progression Chart: The left panel dramatically illustrates the pipeline's journey—stages 1-5 show flat performance hovering at the baseline (~0.51), representing near-chance discrimination. Then stage 6 (hyperparameter tuning) shows a sharp **vertical spike to 0.60**, representing the breakthrough from random to meaningful prediction. This visual alone tells the complete story: feature engineering failed, model optimization succeeded.

Feature Count by Stage: The right panel shows the dimensionality journey: starting at 61 features, attempting aggressive reduction to 30 (filter) and 15 (forward/backward), expanding to 156 (PCA components), then returning to all 61 features with proper tuning. The chart reveals that **feature quantity was never the constraint**—model architecture was.

Cumulative Feature Importance: Both importance curves show identical steep ascent patterns, confirming that just **9 features capture 90% of importance**. This consistency across visualizations validates that the tuned model successfully identified and exploited the concentrated predictive signal while ignoring noise from the remaining 85% of features.

Top 20 Feature Importances: The horizontal bar chart shows ps_car_13's dominance visually—its bar extends 2.4× longer than the second-place feature (ps_ind_05_cat). The rapid drop-off after the top 3 features creates a characteristic "hockey stick" pattern, with features ranked 11-20 barely visible, reinforcing the extreme concentration of predictive power.

Key Findings Synthesis

1. Baseline Challenge (AUROC 0.5086)

- Severe overfitting: Training AUROC 1.00 vs Validation 0.51
- Class imbalance: 94.9% majority vs 5.1% minority
- Near-random discrimination on validation set
- Model memorized training data without learning generalizable patterns

2. Feature Engineering Failure (Stages 2-5)

- All dimensionality reduction attempts failed to improve performance

- Reduction from 61→30→15 features: No benefit (-0.24% to -0.74%)
- PCA transformation to 156 components: Captured 90% variance but maintained chance-level performance (-0.27%)
- Conclusion: **Feature quantity/transformation wasn't the problem; model overfitting was**

3. Hyperparameter Tuning Breakthrough (Stage 6)

- **17.09% improvement** (0.5086 → 0.5956 AUROC)
- Optimal parameters enforced aggressive regularization:
    - max_depth=5: Shallow trees prevent overfitting
    - min_samples_leaf=8: Robust leaf predictions
    - criterion=entropy: Maximizes information gain for minority class
    - class_weight=None: Surprisingly, explicit balancing was counterproductive
- Model complexity: 32 leaves, 63 nodes (highly constrained vs unlimited baseline)

4. Feature Importance Discovery

- **ps_car_13 dominates**: 35.7% of all importance (single feature > 1/3 of predictive power)
- **Top 3 features**: 60.3% importance (ps_car_13 + ps_ind_05_cat + ps_ind_17_bin)
- **Top 10 features**: 93.2% importance
- **44 features (72%)**: Zero contribution—pure noise
- Concentrated signal validates that shallow depth=5 constraint was optimal

5. Model Architecture Insights The tuned model's success stems from **constraint-driven signal extraction**:

- Shallow depth forces selection of only the strongest splits (primarily ps_car_13-based)
- Entropy criterion prioritizes information gain valuable for minority class
- Conservative leaf size (min=8) prevents overfitting to individual minority instances
- Returning to all 61 features allows the shallow tree to select optimal splits from full space

Critical Takeaways

What Worked: ✓ Aggressive regularization through shallow trees (max_depth=5) ✓ Entropy-based splitting for minority class sensitivity ✓ Using all 61 features with strong architectural constraints ✓ Systematic hyperparameter search across 320 combinations

What Didn't Work: ✗ Filter-based feature selection (univariate statistics insufficient) ✗ Wrapper methods (forward/backward selection) ✗ PCA transformation (variance ≠ discriminative power) ✗ Dimensionality reduction without addressing overfitting

Fundamental Lesson: For extremely imbalanced datasets (19:1 ratio) with weak signals, **model architecture and regularization are exponentially more important than feature engineering**. The signal existed in the original 61 features—it simply required proper model constraints to extract without overfitting to noise.

Remaining Limitations & Future Directions

Despite the breakthrough, **AUROC 0.60 represents modest discrimination**:

- Still misses many minority class instances (better than 50/50, but far from reliable)
- Single feature dependence (ps_car_13) creates vulnerability
- Shallow tree limits ability to capture complex interactions

Recommended Next Steps:

1. **Ensemble methods**: Random Forest or XGBoost with calibrated shallow trees
2. **Advanced resampling**: SMOTE/ADASYN to synthetically balance training data
3. **Cost-sensitive learning**: Explicitly weight minority class errors higher
4. **Feature interactions**: Engineer ps_car_13 × ps_ind_05_cat cross-terms
5. **Anomaly detection**: Treat minority class as anomaly detection problem
6. **Deep learning**: Neural networks with class-weighted loss functions

The pipeline successfully diagnosed the problem (overfitting, not features) and achieved meaningful improvement through principled regularization. The foundation is now established for advanced techniques to push beyond AUROC 0.60 toward production-ready performance.

Conclusion

This comprehensive study on the Enhanced Safe Driver Prediction Challenge successfully identified and addressed the fundamental obstacles in predicting insurance claim risk from severely imbalanced data. Through systematic experimentation across multiple methodologies, the project achieved a **17.09% improvement in AUROC** (from 0.5086 to 0.5956), representing a crucial transition from chance-level prediction to meaningful discriminative capability.

Key Findings

The project's most significant discovery challenges conventional machine learning wisdom: **feature engineering alone cannot overcome severe model overfitting**. Despite rigorous statistical testing and dimensionality reduction attempts—including filter selection (30 features), forward/backward elimination (15 features), and PCA transformation (156 components)—all approaches maintained near-chance performance (AUROC 0.505-0.507). The breakthrough came exclusively from **hyperparameter optimization**, which enforced aggressive regularization through shallow trees (max_depth=5), entropy-based splitting, and conservative leaf size (min_samples_leaf=8).

Counterintuitively, the optimal model used **all 61 original features without class balancing** (class_weight=None), outperforming carefully selected feature subsets. This demonstrates that proper

architectural constraints allow models to naturally identify important features while ignoring noise, rather than requiring explicit feature reduction. The analysis revealed extreme feature concentration: `ps_car_13` alone accounts for 35.7% of predictive power, with just 9 features capturing 90% of importance, while 44 features (72%) contribute nothing—validating the shallow tree's efficiency in signal extraction.

Methodological Contributions

The systematic six-stage pipeline provides a replicable diagnostic framework for imbalanced classification problems. The consistent failure of feature engineering methods (stages 2-5) followed by hyperparameter tuning success (stage 6) establishes a clear pattern: when dimensionality reduction fails to improve chance-level performance, **model overfitting—not feature quality—is the root cause**. This insight, often overlooked in literature emphasizing successful techniques, offers valuable guidance for practitioners facing similar challenges.

Practical Implications and Limitations

From an industry perspective, AUROC 0.60 represents meaningful but incomplete progress. The model now ranks high-risk drivers better than random assignment and successfully addresses the baseline's catastrophic minority class failure (8% recall → improved discrimination). However, the modest performance level and heavy dependence on a single feature (`ps_car_13`) indicate this solution is a **foundation rather than a production-ready system**. The 19:1 class imbalance and weak individual feature signals suggest that insurance claim prediction remains inherently challenging due to the stochastic nature of accidents and human behavior.

Future Directions

The established baseline creates opportunities for advanced techniques: ensemble methods (Random Forest, XGBoost) with calibrated shallow learners, SMOTE/ADASYN resampling, deep learning with class-weighted losses, feature interaction engineering (particularly `ps_car_13 × ps_ind_05_cat`), and anomaly detection frameworks treating the minority class as outliers. These approaches could potentially push performance beyond AUROC 0.60 toward production-viable levels (0.70-0.80+).

Final Reflection

The project's ultimate lesson is that **understanding why methods fail is as valuable as knowing which succeed**. The multiple unsuccessful feature engineering attempts provided diagnostic insight that model architecture, not feature selection, was the critical constraint. For extremely imbalanced datasets with weak signals, constraint-driven learning through shallow architectures forces identification of only the most robust, generalizable patterns—a principle with broad applicability beyond insurance to fraud detection, medical diagnosis, and rare event prediction domains. The journey from perfect training performance (AUROC 1.00) but chance validation (0.51) to balanced modest performance (0.60) exemplifies how principled regularization extracts genuine signal from noise, establishing a methodological foundation for advancing predictive analytics in challenging real-world contexts.

Thank you!