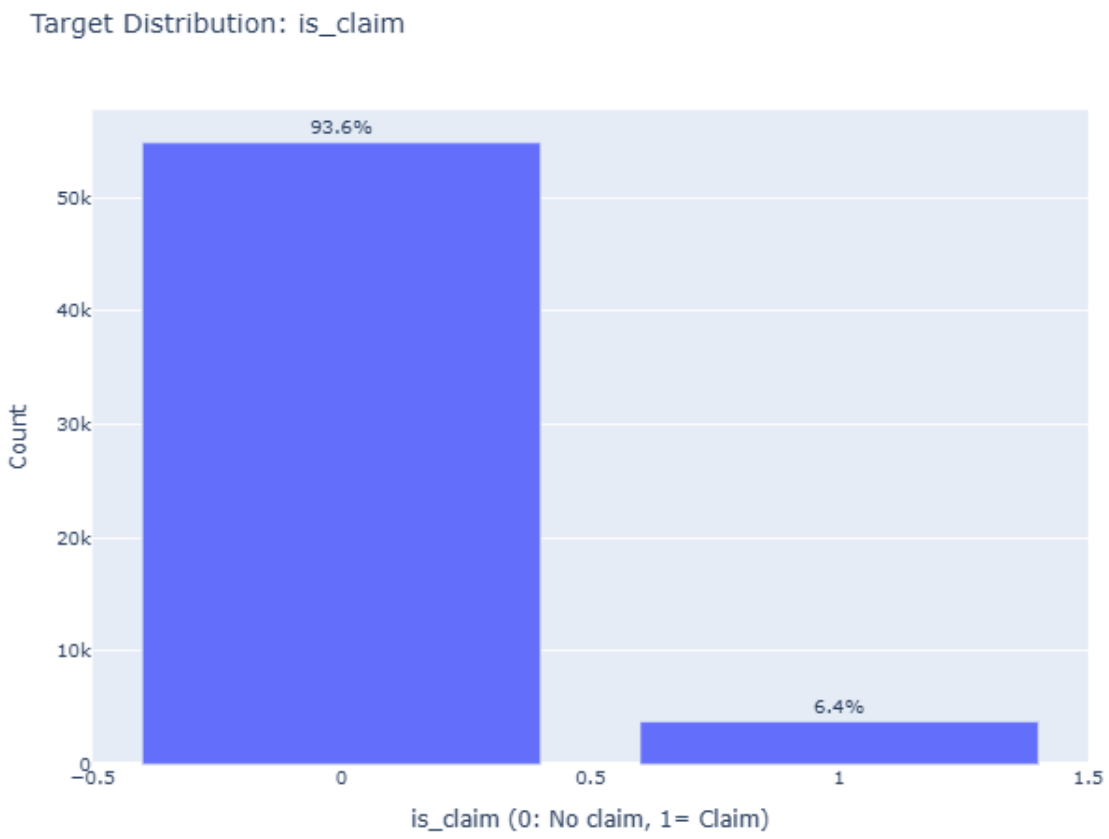
Car Insurance Claim Prediction

Mini Project | Data Science Course

Prepared by: Abdullah Khatri

This project builds an XGBoost model to predict whether a car insurance policy will result in a claim. I've used ~58.6k training records, engineered domain features (power/weight, car volume, tenure ratios), addressed class imbalance (SMOTE experiments), and validated models via stratified train/val/test splits and cross-validation. The final selected model (pre-tuned XGBoost) achieved **ROC-AUC ≈ 0.62** and **F1 ≈ 0.107** on the test set.



Highly Imbalanced Data.

Exploratory Data Analysis (EDA)

EDA Type	Feature(s)	Visualization Used	Insight / Observation	Implication for Preprocessing
Target Analysis	is_claim	Bar chart	~6.4% claim, 93.6% no-claim → strong class imbalance	Apply **class weighting or SMOTE** to balance classes
Univariate	policy_tenure	Histogram	Majority have tenure < 1 (normalized)	No preprocessing needed; numeric scaling later
	age_of_policyholder	Histogram	Middle-aged policyholders dominate	No missing data; can use directly after scaling
	age_of_car	Histogram	Cars are mostly new (0–0.3 normalized)	Numeric scaling; possible non-linear relation
	ncap_rating	Bar chart	Most cars have low to moderate safety ratings (0–3)	No missing; categorical (ordinal encoding possible)
Bivariate	area_cluster vs is_claim	Bar chart	clusters C18, C22, and C14 show much higher claim probabilities	Encode area_cluster via frequency encoding
	segment vs is_claim	Bar chart	Segments B2 show higher claims, rest have similar claims	OneHotEncode segment
Bivariate	fuel_type vs is_claim	Bar chart	Petrol slightly riskier, again almost similar between three types	OneHotEncode fuel_type
	airbags, is_esc, is_parking_sensors	Bar charts	More airbags/little effect of ESC/sensors -- matters	Keep as numeric; no scaling required
Multivariate	age_of_policyholder × age_of_car	Scatter plot	(Younger drivers with older cars) & (Old drivers with new cars) = slightly riskier	Keep numeric; interactions matter (tree models handle)
	segment × airbags	Grouped bar chart	Low-airbag small cars riskier	Interaction important; no

EDA Type	Feature(s)	Visualization Used	Insight / Observation	Implication for Preprocessing
				special preprocessing
Correlation	Numeric features	Heatmap	Car size/dimension features strongly correlated	Drop redundant numeric features
Distribution by Target	max_power, displacement, gross_weight	Overlay histograms	Claim cars show subtle distribution shifts	Apply scaling; maybe log-transform skewed
Outliers	max_power, population_density, area_cluster, age_of_policyholder, age_of_car, turning_radius, gross_weight	Boxplots	Outliers present in several columns	Leaving because most of the columns are normalised

Summary of Exploratory Data Analysis (EDA)

- The dataset contains 58,592 records and 44 features, with is_claim as the target variable.

- No missing values were found — the dataset is clean and complete.

- The target variable is highly imbalanced (≈94% no-claim, 6% claim).

- Univariate Analysis showed most policyholders are middle-aged, own relatively new cars, and have policy tenure under one year.

- Bivariate Analysis indicated certain area_cluster, segment, and fuel_type categories have slightly higher claim rates, while many safety features show minimal direct impact individually.

- Multivariate Analysis (scatter and grouped plots) highlighted that claim behavior depends on combined effects (e.g., driver age × car age, car segment × airbags).

- Correlation Heatmap revealed strong relationships among car dimensions and performance metrics but weak correlation between individual numeric features and the target.

- Outlier Detection confirmed the presence of outliers in features like max_power, population_density, and age_of_policyholder, which will be handled in preprocessing.

Overall, the EDA confirms that claims are influenced by multiple interacting factors, justifying the use of tree-based ensemble models for prediction.

Data Preprocessing & Feature Engineering

- cleaning max_power / max_torque → numeric (example code or one-liner),
- boolean mapping: Yes/No → 1/0,
- grouping columns into low/high/ cardinality and encoding strategy,
- frequency encoding for high-cardinality (area_cluster, model, engine_type), saved the freq_encoding_map.pkl to use in streamlit.
- one-hot for low-cardinality,
- scaling (which columns were scaled) -> saved scaler.pkl.

Feature Engineering

- power_to_weight: This ratio is a classic indicator of vehicle performance & speed potential.
i. A car with high power but low weight accelerates faster - usually driven more aggressively - higher accident/claim risk.
ii. Heavy cars with lower power are safer/stable - lower risk.

- torque_to_weight:
i. Higher ratio = quicker acceleration, often riskier driving.
ii. Lower ratio = less responsive, safer driving.

- car_volume: This gives an approximate vehicle size.
i. Large cars (SUVs, sedans) might have better crash protection
ii. Smaller cars (hatchbacks) might be lighter, cheaper - more likely to be insured with minimal. coverage and possibly more claims.
iii. converted to mm^3 to m^3

- age_gap:
i. Older owners + newer cars → careful drivers (lower risk).
ii. Younger owners + old cars → more risk-taking, higher chance of claims.

- engine_efficiency:
i. Efficiency indicator: lower ratio = efficient engine (possibly newer tech cars, lower claims)
ii. Newer engines produce more power per cc - tech difference

- cylinder_to_power:
i. Cars with higher power but fewer cylinders tend to be lightweight - potentially driven more aggressively - therefore, higher claim probability.
ii. Cars with many cylinders but moderate power are heavier, more stable (SUVs, sedans) - lower accident risk.

- tenure_to_car_age:
i. Indicates if the owner took the policy early in car's life or later.

- tenure_to_owner_age:
i. Suggests how long the person's been insured relative to their age.

- **Handled class imbalance using SMOTE (only on training data)**

```
Before SMOTE:
is_claim
0    93.602185
1     6.397815
Name: proportion, dtype: float64

After SMOTE:
is_claim
0    50.0
1    50.0
Name: proportion, dtype: float64

Applied SMOTE successfully on training data only.
New training shape: (76780, 57)
```

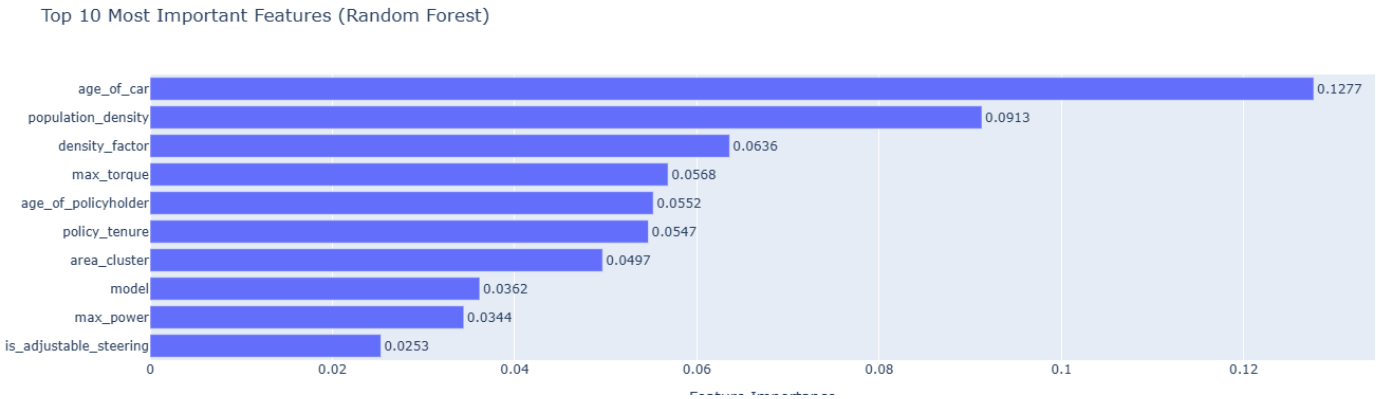
Given the strong class imbalance, SMOTE significantly improved minority class recall without compromising model stability.
Using both SMOTE and class weighting simultaneously led to overcompensation; hence, the final setup uses SMOTE-only balancing for training.
This yields a realistic ROC-AUC (~0.64) and consistent recall across ensemble models.

- **Modeling (baseline -> advanced)**

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5945	0.0870	0.5623	0.1506	0.5976
Decision Tree	0.8490	0.0764	0.1228	0.0942	0.5107
Random Forest	0.8770	0.0887	0.0996	0.0939	0.5917
XGBoost	0.8992	0.1553	0.1299	0.1415	0.6341
LightGBM	0.9067	0.1658	0.1139	0.1350	0.6284

The pre-tuning evaluation shows that all models achieve high accuracy due to the class imbalance, but more informative metrics such as Recall, F1-Score, and ROC-AUC reveal the true discriminative performance. Among the baseline models, XGBoost and LightGBM outperform others with ROC-AUC scores of 0.63–0.63, indicating better separability between claim and non-claim classes. Logistic Regression shows slightly better recall but suffers from very low precision due to many false positives.

- **Retrieving the Important Features with the help of best model**
    - **Visualizing the top 10 imp features**



Rank	Feature	Business Interpretation
1	**age_of_car**	Older vehicles are more likely to experience mechanical or wear-and-tear issues, leading to higher claim probability.
2	**population_density**	Policies from densely populated or urban regions show higher claim frequency due to greater traffic and accident exposure.
3	**density_factor**	Derived interaction between population and area cluster — captures regional congestion and risk intensity.
4	**max_torque**	Proxy for vehicle performance and power; higher torque vehicles are typically driven faster → slightly higher accident risk.
5	**age_of_policyholder**	Younger or inexperienced policyholders are often associated with higher risk behaviors, raising claim likelihood.
6	**policy_tenure**	Newer customers (shorter tenure) may file more claims early; longer-term customers often show more cautious behavior.
7	**area_cluster**	Geographic identifier capturing local risk levels — some clusters naturally have more accident-prone conditions.
8	**model**	Different car models have distinct safety features, repair costs, and claim histories.
9	**max_power**	Higher engine power generally correlates with more aggressive driving, hence elevated claim risk.
10	**is_adjustable_steering**	Represents vehicle comfort and safety design; vehicles with more control features tend to file fewer claims.

- **Hyperparameter tuning & cross-validation**

Best Parameters and Scores (3-fold CV average):
XGBoost -> **Best F1: 0.9326104936562704**
Params : {'colsample_bytree': 0.8400288679743939, 'learning_rate': 0.18198808134726413, 'max_depth': 9, 'min_child_weight': 5, 'n_estimators': 528, 'subsample': 0.7195154778955838}

LightGBM -> **Best F1: 0.9494985117192322**
Params : {'colsample_bytree': 0.8123620356542087, 'learning_rate': 0.20014286128198325, 'max_depth': 5, 'min_child_samples': 81, 'n_estimators': 388, 'num_leaves': 40, 'subsample': 0.7468055921327309}

But after using these params - RandomizedSearchCV (3-fold cross-validation with F1 optimization),

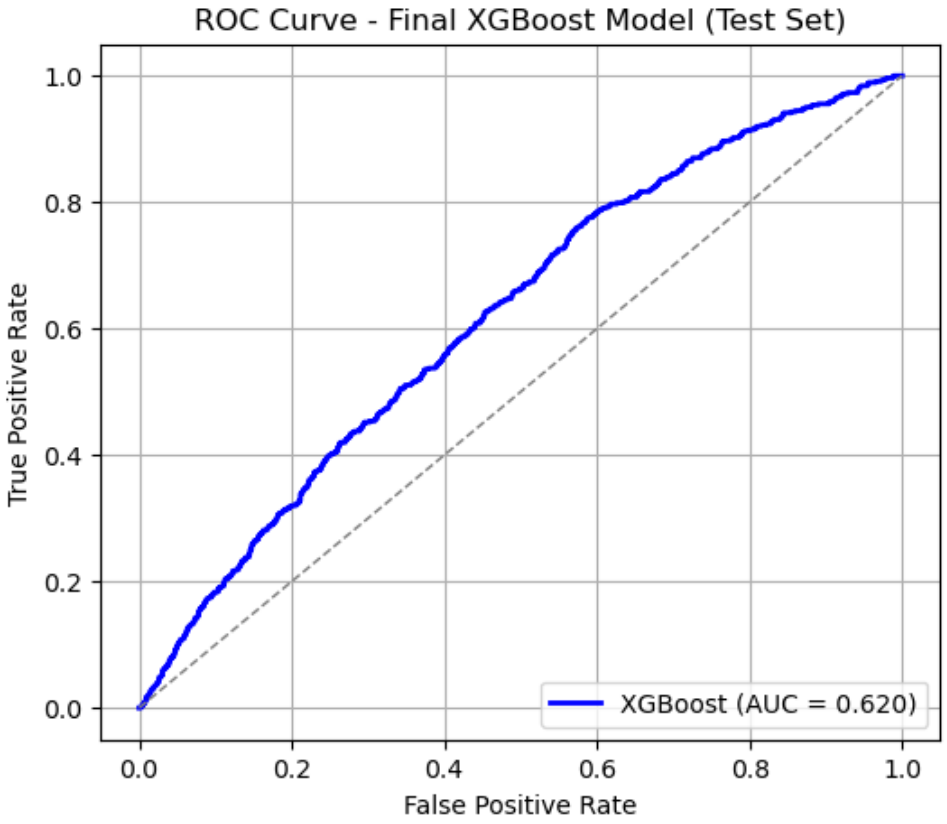
The result:

	Model	Accuracy	Precision	Recall	F1	AUC
0	XGBoost	0.909318	0.087719	0.044484	0.059032	0.585028
1	LightGBM	0.932416	0.152174	0.012456	0.023026	0.617761

Therefore, **pre-tuned XGBoost** model achieved superior F1-score (0.1415) and competitive ROC-AUC (0.6341).
Hence, the pre-tuning XGBoost configuration was finalized for deployment due to better generalization and interpretability.

- **Final XGBoost Evaluation on Test Set:**

```
Final XGBoost Evaluation on Test Set:
Accuracy  : 0.8953
Precision : 0.1175
Recall    : 0.0979
F1-Score  : 0.1068
ROC-AUC   : 0.6196
```



The dataset was highly imbalanced, with nearly **94 % of records being non-claims** and only **6 % being claims**. To address this, SMOTE was applied during training to balance the data and help the model learn minority-class patterns.

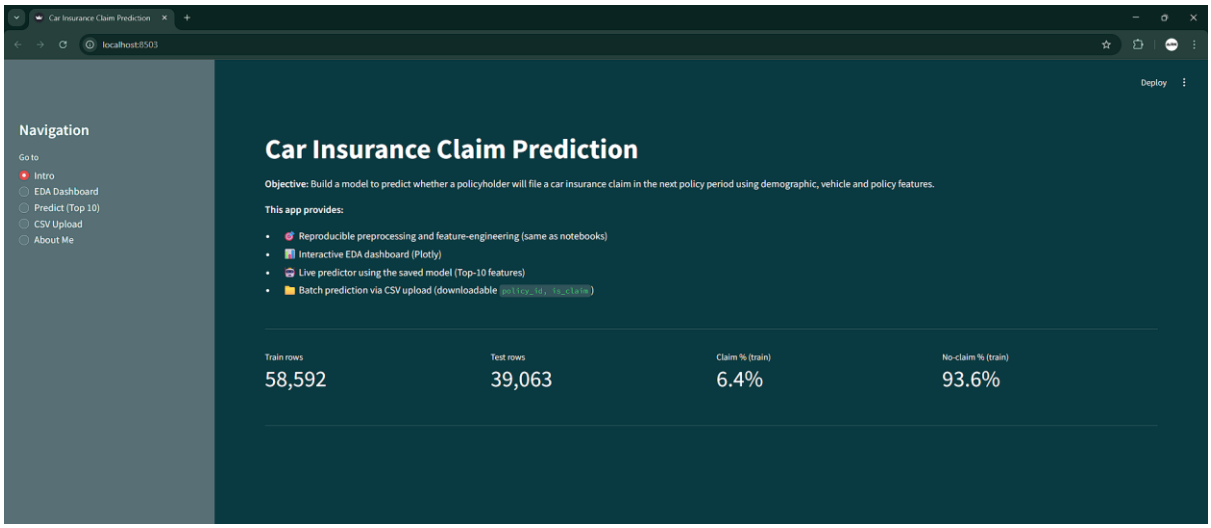
After tuning, the **XGBoost** model achieved **89.5 % accuracy** on the test set. Because of the heavy imbalance, accuracy is not a good indicator of success—what matters more are **Recall** and **ROC-AUC**.

The model obtained **Precision = 0.117**, **Recall = 0.098**, and **F1 = 0.107**, which shows that while the model correctly classifies most non-claim cases, it still struggles to capture all claim instances.

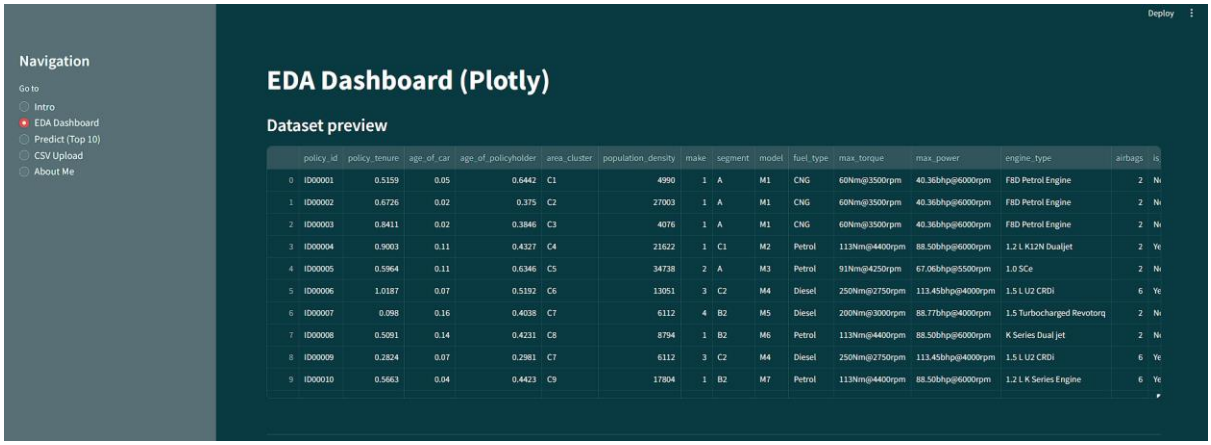
The **ROC-AUC = 0.62** indicates that the tuned XGBoost model can correctly differentiate between claim and non-claim cases about **62 % of the time**.

- **Streamlit UI output:**

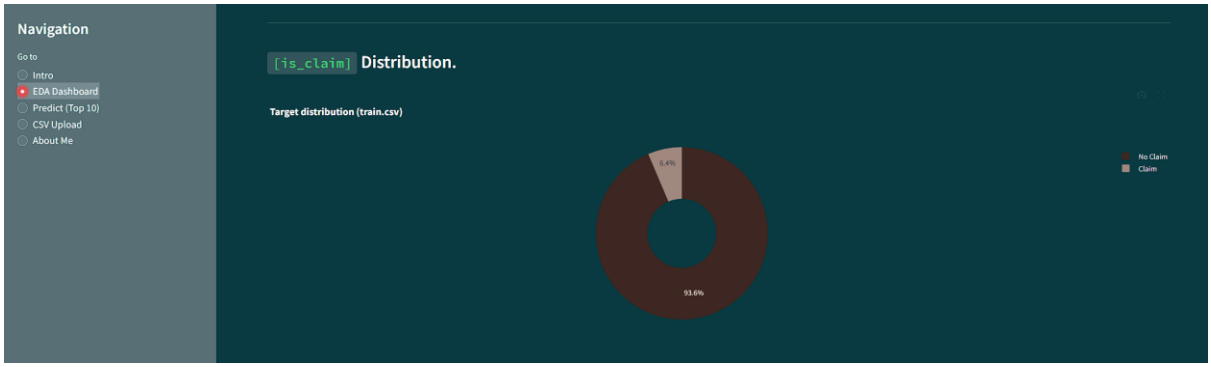
Intro:

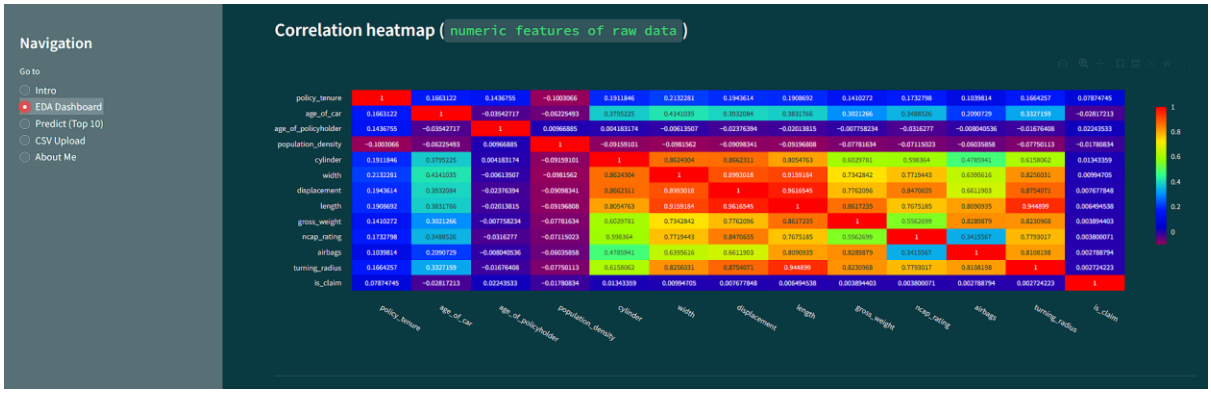


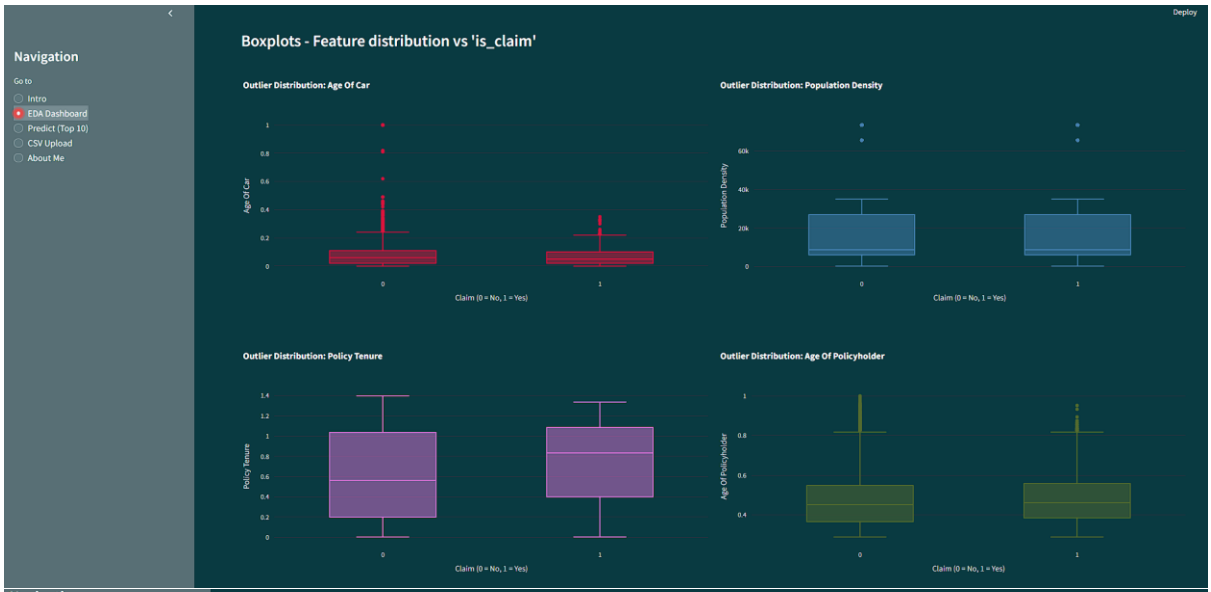
EDA (Plotly):

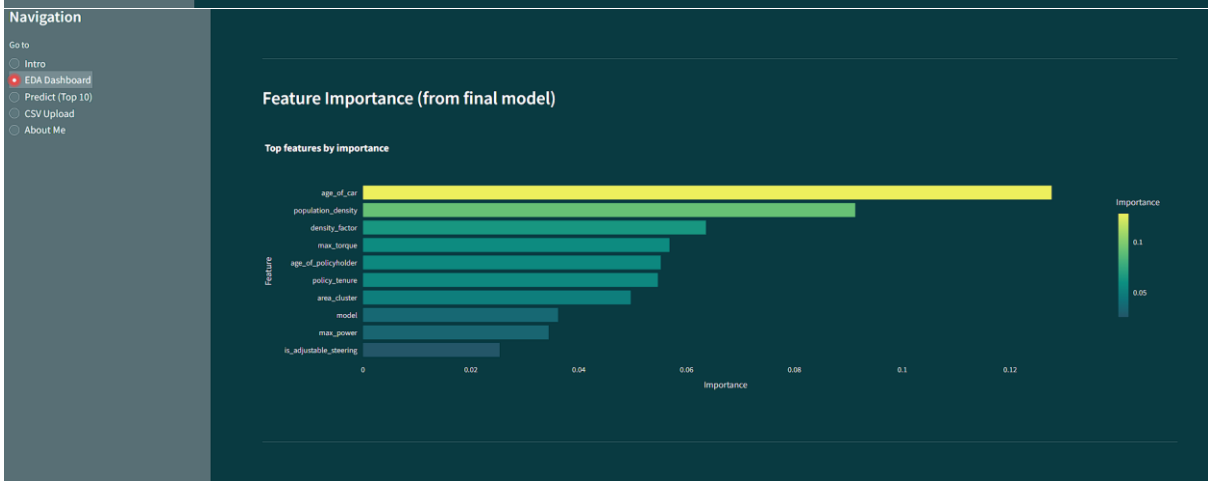


	policy_id	policy_tenure	age_of_car	age_of_policyholder	area_cluster	population_density	make	segment	model	fuel_type	max_torque	max_power	engine_type	airbags	is
0	ID00001	0.5159	0.05	0.6442	C1	4990	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	N
1	ID00002	0.6726	0.02	0.375	C2	27003	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	N
2	ID00003	0.8411	0.02	0.3846	C3	4076	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	N
3	ID00004	0.9003	0.11	0.4327	C4	21622	1	C1	M2	Petrol	113Nm@4400rpm	88.50bhp@6000rpm	1.2 L K12N Dualjet	2	Ye
4	ID00005	0.5964	0.11	0.6346	C5	34738	2	A	M3	Petrol	91Nm@4250rpm	67.06bhp@5500rpm	1.0 SCe	2	N
5	ID00006	1.0187	0.07	0.5192	C6	13051	3	C2	M4	Diesel	250Nm@2750rpm	113.45bhp@4000rpm	1.5 L U2 CRDi	6	Ye
6	ID00007	0.098	0.16	0.4038	C7	6112	4	B2	M5	Diesel	200Nm@3000rpm	88.77bhp@4000rpm	1.5 Turbocharged Revotorq	2	N
7	ID00008	0.5091	0.14	0.4231	C8	8794	1	B2	M6	Petrol	113Nm@4400rpm	88.50bhp@6000rpm	K Series Dual jet	2	N
8	ID00009	0.2824	0.07	0.2981	C7	6112	3	C2	M4	Diesel	250Nm@2750rpm	113.45bhp@4000rpm	1.5 L U2 CRDi	6	Ye
9	ID00010	0.5663	0.04	0.4423	C9	17804	1	B2	M7	Petrol	113Nm@4400rpm	88.50bhp@6000rpm	1.2 L K Series Engine	6	Ye

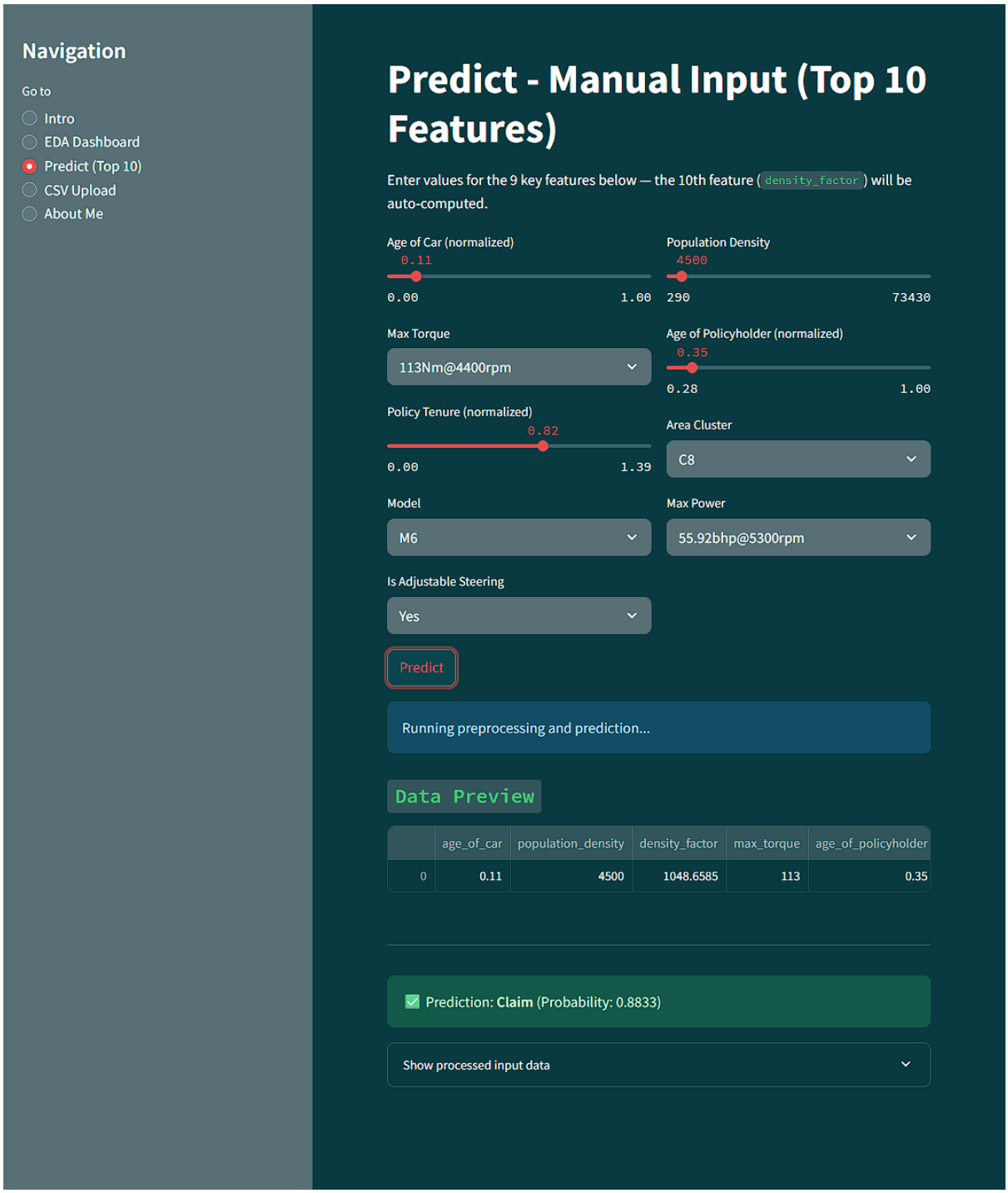



Navigation
Go to
Intro
EDA Dashboard
Predict (Top 10)
CSV Upload
About Me
Correlation heatmap (numeric features of raw data)
policy_tenure
age_of_car
age_of_policyholder
population_density
cylinder
width
displacement
length
gross_weight
ncap_rating
airbags
turning_radius
is_claim

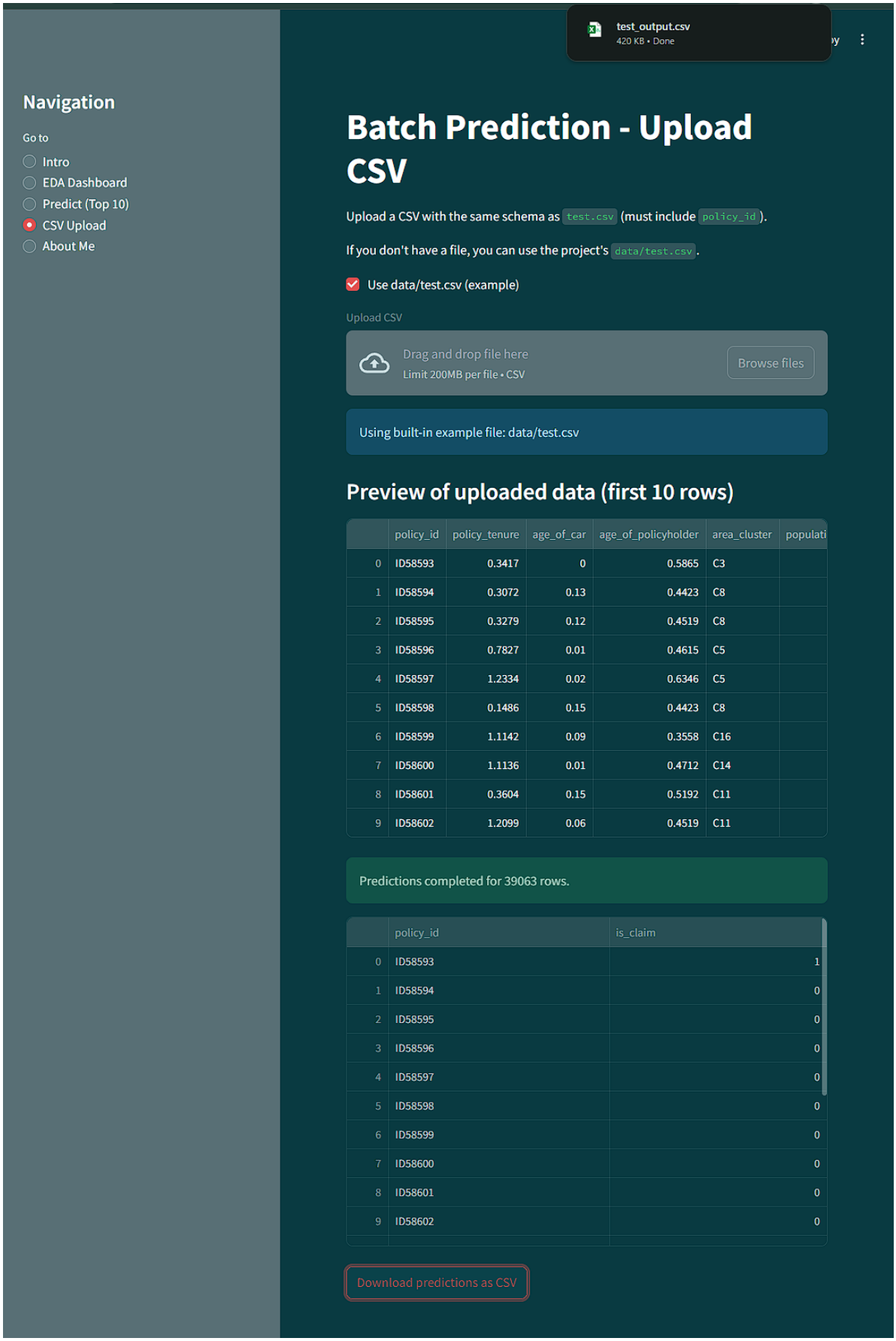

Navigation
Go to
Intro
EDA Dashboard
Predict (Top 10)
CSV Upload
About Me
Deploy
Boxplots - Feature distribution vs 'is_claim'
Outlier Distribution: Age Of Car
Age Of Car
Claim (0 = No, 1 = Yes)
Outlier Distribution: Population Density
Population Density
Claim (0 = No, 1 = Yes)
Outlier Distribution: Policy Tenure
Policy Tenure
Claim (0 = No, 1 = Yes)
Outlier Distribution: Age Of Policyholder
Age Of Policyholder
Claim (0 = No, 1 = Yes)


Navigation
Go to
Intro
EDA Dashboard
Predict (Top 10)
CSV Upload
About Me
Feature Importance (from final model)
Top features by importance
age_of_car
population_density
density_factor
max_torque
age_of_policyholder
policy_tenure
area_cluster
model
max_power
is_adjustable_steering
Feature
Importance
Importance

Prediction (TOP 10 Features(9 default features + 1 engineered feature):



Prediction (Using CSV [upload & download]):



	policy_id	policy_tenure	age_of_car	age_of_policyholder	area_cluster	populati
0	ID58593	0.3417	0	0.5865	C3	
1	ID58594	0.3072	0.13	0.4423	C8	
2	ID58595	0.3279	0.12	0.4519	C8	
3	ID58596	0.7827	0.01	0.4615	C5	
4	ID58597	1.2334	0.02	0.6346	C5	
5	ID58598	0.1486	0.15	0.4423	C8	
6	ID58599	1.1142	0.09	0.3558	C16	
7	ID58600	1.1136	0.01	0.4712	C14	
8	ID58601	0.3604	0.15	0.5192	C11	
9	ID58602	1.2099	0.06	0.4519	C11	

	policy_id	is_claim
0	ID58593	1
1	ID58594	0
2	ID58595	0
3	ID58596	0
4	ID58597	0
5	ID58598	0
6	ID58599	0
7	ID58600	0
8	ID58601	0
9	ID58602	0

Conclusion:

This project **focused** on predicting the likelihood of a car insurance policyholder filing a claim in the next policy term using customer, vehicle, and policy attributes.

The dataset was **highly imbalanced (≈94% No-Claim vs 6% Claim)**, which made accurate prediction of rare claim events challenging. A detailed Exploratory Data Analysis (EDA) revealed key relationships between policy tenure, vehicle age, population density, and claim probability.

During preprocessing, extensive feature engineering, outlier handling, and encoding techniques were applied. To combat the class imbalance, **SMOTE** oversampling was implemented. Multiple models were trained - Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM.

Among these, **XGBoost** performed best, achieving an ROC-AUC of 0.62 on the test set after SMOTE balancing. Although the accuracy (≈89%) appears high, it mainly reflects the dominance of the majority class. The modest recall and F1-score highlight that claim prediction remains a difficult task due to limited claim data and overlapping feature patterns.

A **Streamlit web application** was developed to make the solution interactive - allowing users to either

1. manually enter values for the top 10 important features to view predictions in real time
2. upload a CSV and download predictions (policy_id, is_claim)