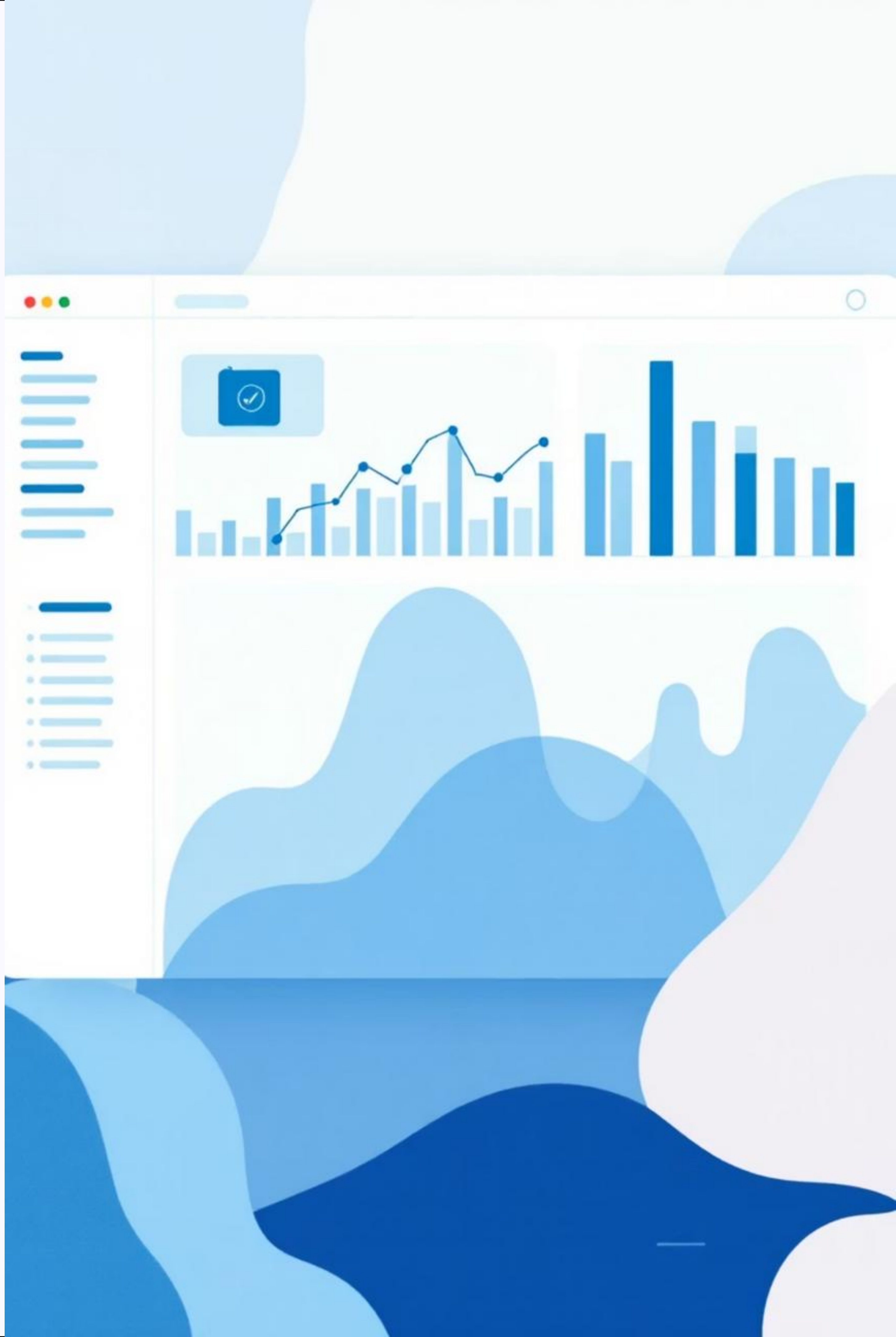
Insurance Customer Response Prediction

Predicting Customer Interest in Insurance Policy Offers



Introduction: Why Predict Customer Response?

This project uses machine learning to predict customer interest in insurance policy offers. By understanding key customer behavior patterns, the solution helps insurance companies target the right customers, improve marketing efficiency, and maximize conversion rates. The model integrates data insights, algorithm evaluation, and a deployed Streamlit application for real-time predictions.

Cost Efficiency

Insurance companies face costly, inefficient marketing without targeted outreach strategies

Conversion Optimization

Predicting customer interest helps optimize campaigns and increase conversion rates dramatically

Our Focus

Predicting if customers will respond positively to vehicle insurance policy offers

Workflow Overview

A systematic approach to building our prediction model

Data Collection & Understanding

Gathering comprehensive customer data and understanding feature relationships

Exploratory Data Analysis

Uncovering patterns, correlations, and insights within the dataset

Data Preprocessing & Feature Engineering

Cleaning data and creating meaningful features for model training

Model Selection & Training

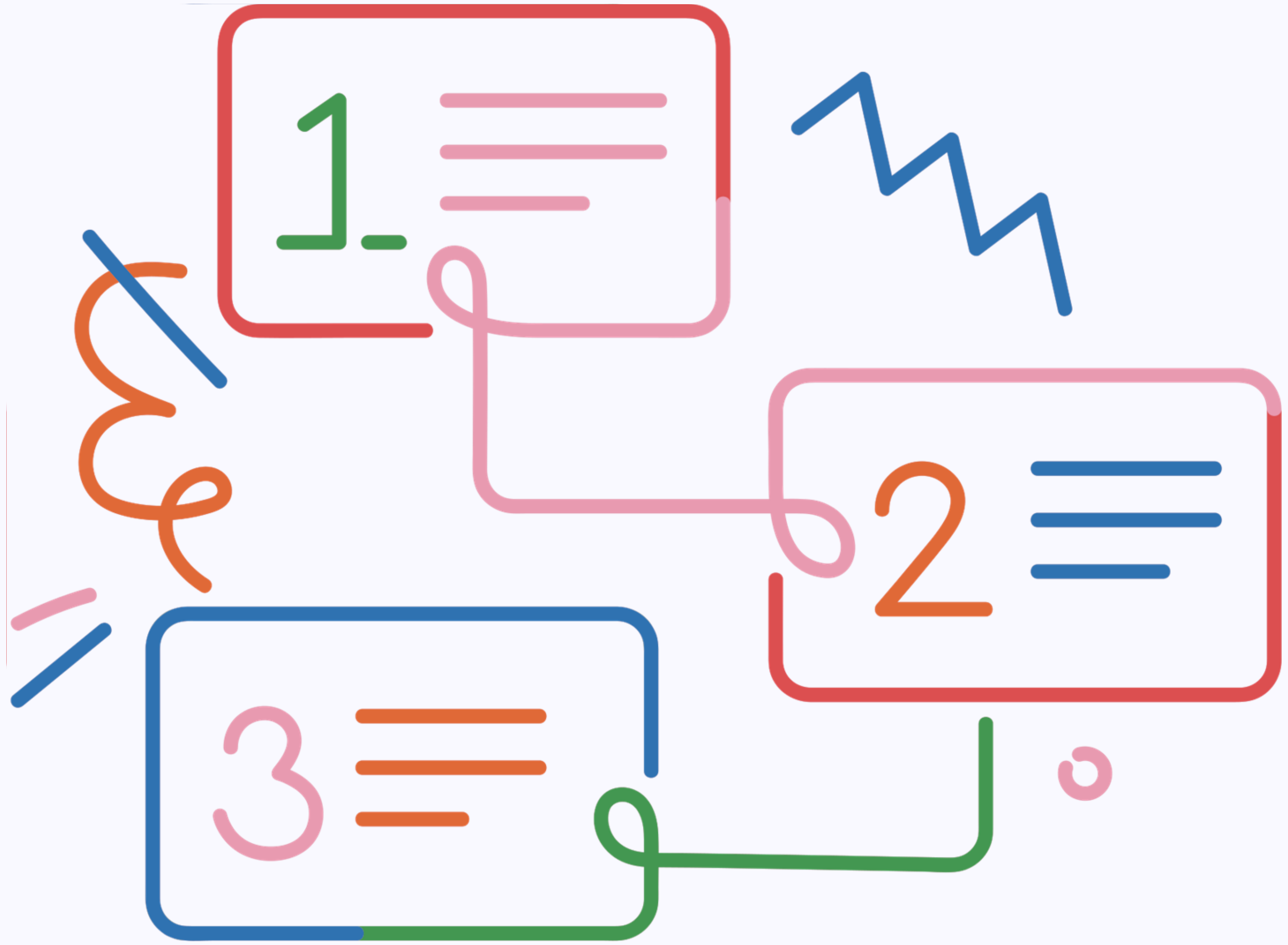
Testing multiple algorithms to find the optimal prediction approach

Model Evaluation & Comparison

Measuring performance metrics and comparing model effectiveness

Deployment via UI

Creating an accessible interface for real-world application



About the Data

Dataset Overview

Our comprehensive dataset contains **381,109 customer records** with over 12 distinct features, providing a robust foundation for predictive modeling.

Key Features

- **Demographics:** Age, Gender, Region_Code
- **Vehicle Information:** Vehicle_Age, Vehicle_Damage
- **Policy Details:** Previously_Insured, Annual_Premium
- **Engagement:** Policy_Sales_Channel, Vintage (days as customer)
- **Others** : Driving_Licence

Target Variable: Response (1 = Interested, 0 = Not Interested)

381K

Customer Records

12

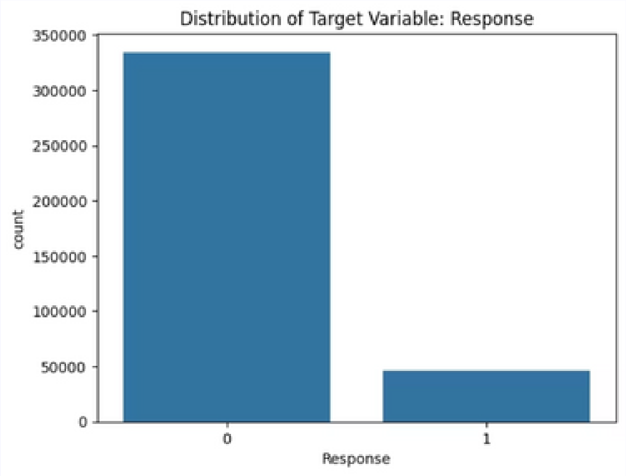
Features

0

Missing Values

Clean, ready-to-use data

Exploratory Data Analysis Highlights

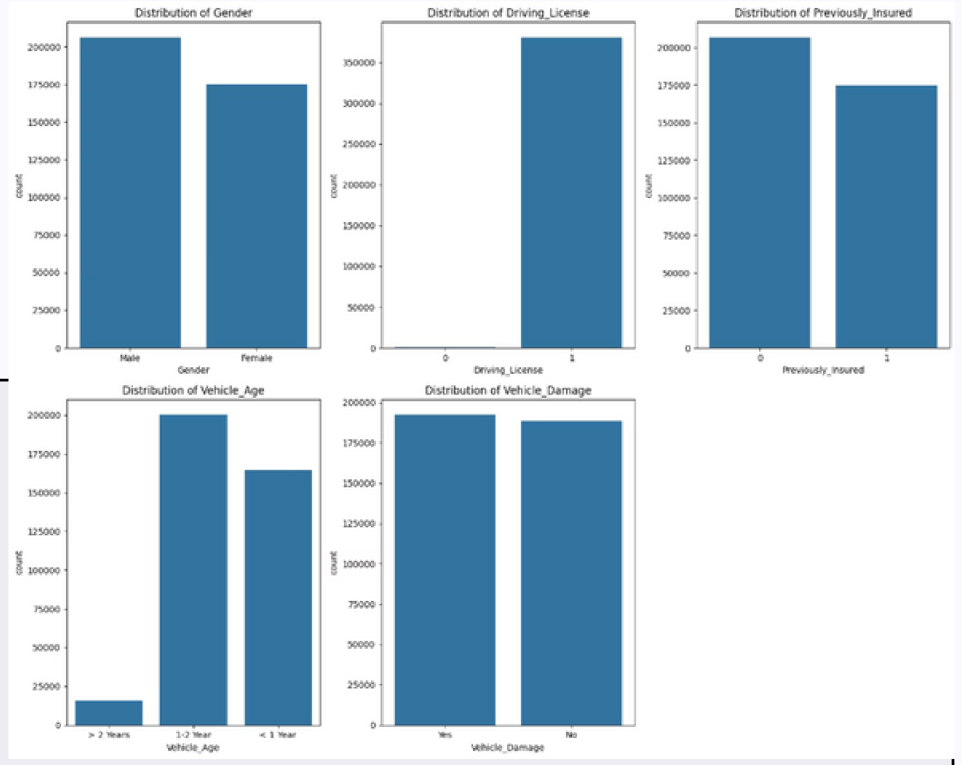


Target Variable is highly imbalanced

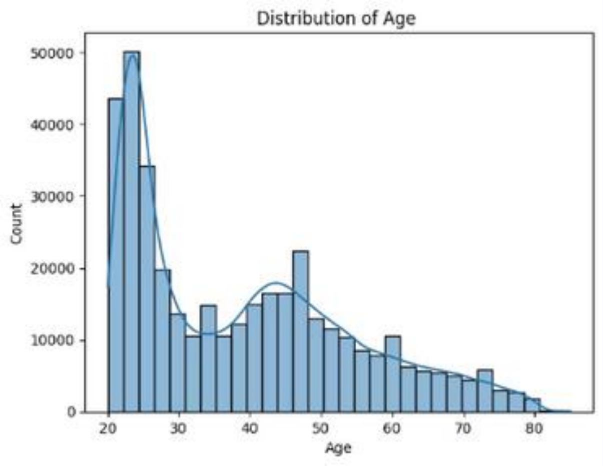
```
Almost 88% do not response positively.
```

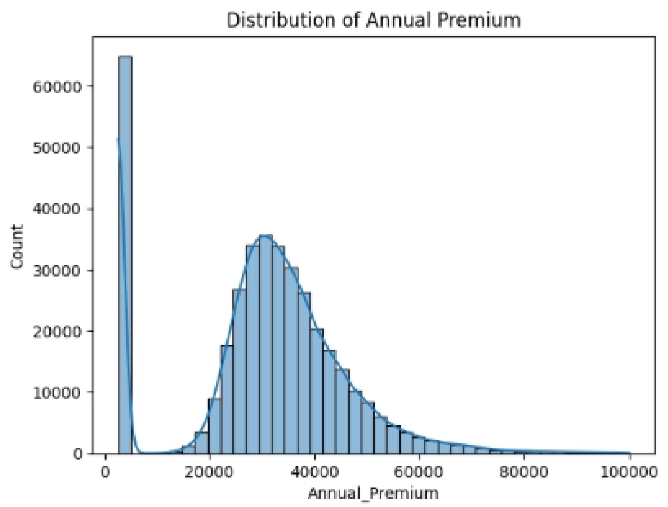
```
- and 12% responded positvely.
```

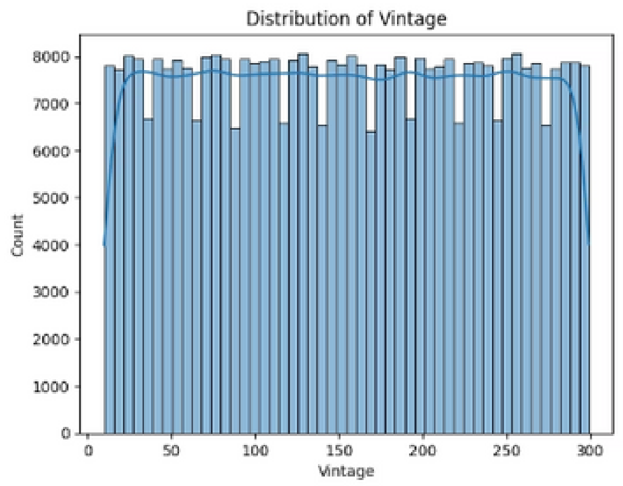
UNIVARIATE ANALYSIS :

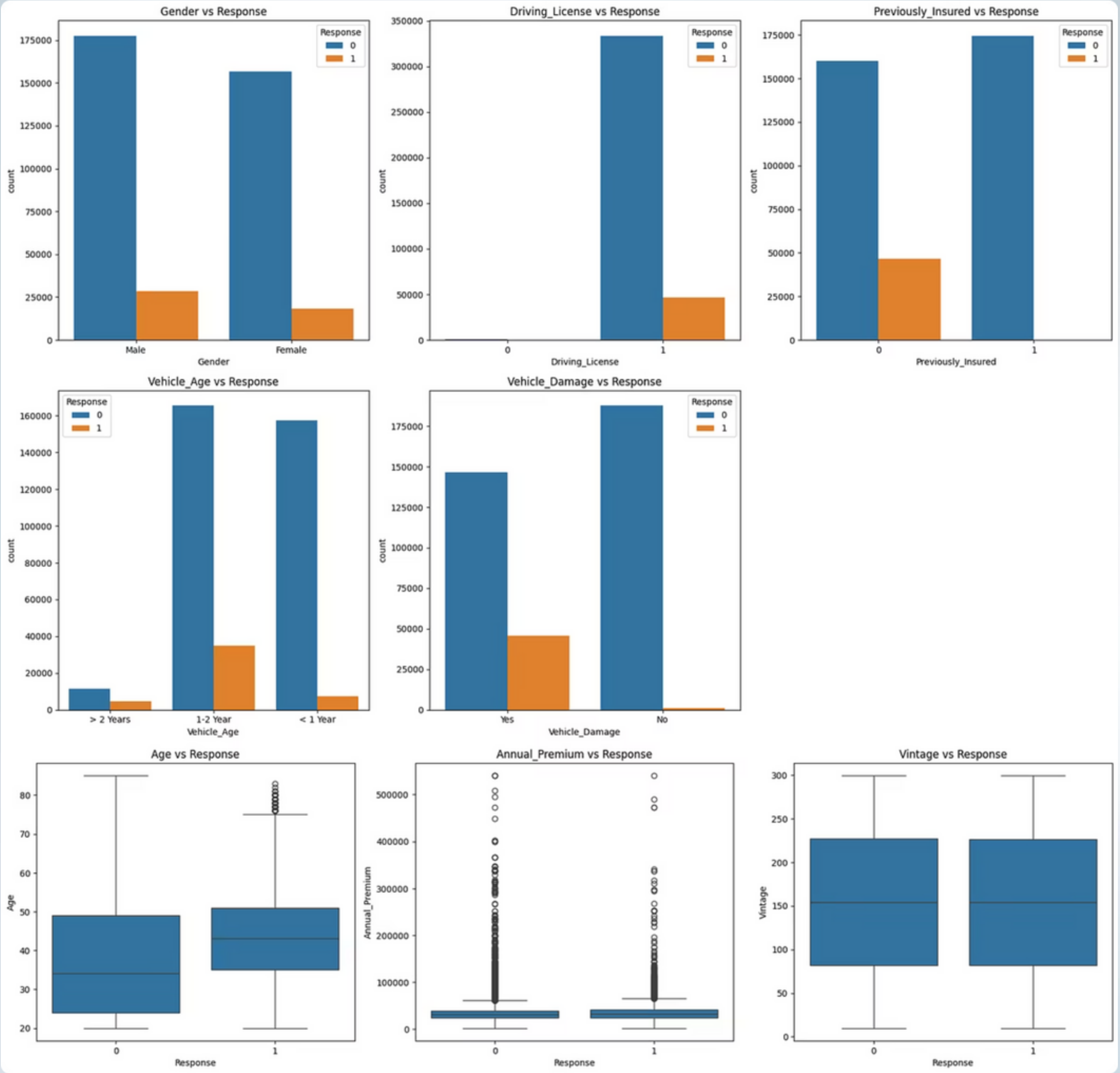


Numeric:
Age: Mostly between 20–50 years; right-skewed with more young customers (20–35).
Annual Premium: Highly right-skewed; majority fall in the ₹20K–50K range.
Vintage: Nearly uniform distribution across 0–300 days, indicating varied customer association duration.
Categorical:
Gender: Slightly more Male customers, but overall balanced.
Driving License: Most customers have DL = 1, making it a low-impact feature.
Previously Insured: Majority are not previously insured—a strong predictor since previously insured customers show lower interest in new policies.
Vehicle Age: Most vehicles are 1–2 years old; few are >2 years. Distinct categories may reflect different renewal/interest behavior.









BIVARIATE ANALYSIS :

- **Gender** does not significantly influence customer interest.
- **Driving License** has almost no variability → contributes minimally to prediction.
- **Previously Insured:**
  - □ Customers **not previously insured (0)** show a **very high response rate**.
  - □ Customers **previously insured (1)** almost never respond.
- **Vehicle Age:**
  - □ **> 2 Years** → highest interest
  - □ **1–2 Years** → moderate interest
  - □ **< 1 Year** → lowest interest
- **Vehicle Damage:**
  - □ Customers with **Vehicle Damage = Yes** respond **much more** → highly important feature.
- **Age:** Responders tend to be slightly older, though the difference is small.
- **Annual Premium:** Distribution is similar for responders and non-responders → **not strongly predictive**.
- **Vintage:** Customer tenure shows **no meaningful effect** on response behavior.

Data Preprocessing

- **Dropped the id column**
  - It is only an identifier
  - Provides no predictive value
  - Helps reduce noise in the dataset
- **Encoded Categorical Features**

Feature	Original Format	Encoding Applied
Gender	Male/ Female	Male = 0, Female = 1
Vehicle Age	< 1 Year, 1–2 Year, > 2 Years	<1 → 0, 1–2 → 1, >2 → 2
Vehicle Damage	Yes / No	Yes = 1, No = 0

- **TRAIN TEST SPLIT**

→ We divided the data into training (80%) and testing (20%) sets.

→ Setting a random state ensures consistent results and using stratify=y maintains a proportional distribution of the target variable in both sets.

- **SPLITING THE DATA INTO x & y**

→ We divided the dataset into two parts: x and y.

→ "x" typically represents the independent Variables, and "y" represents the Dependent (target variable) that we want to predict or understand.

Feature Scaling (Standardization)

Why Scaling Was Needed

- Numerical features (Age, Annual Premium, Vintage) are on **different scales**→ e.g., Age ~30, Premium ~30,000

- Models like Logistic Regression and even tree ensembles perform better when features are **scaled**
- Ensures **fair contribution** of each feature to the model.
- What Was Done Description :

Step	Description
Selected numeric columns	Age, Annual Premium, Vintage
Fitted StandardScaler on training data only	Prevents data leakage
Transformed training data	Converts values to standardized scale
Applied same scaler to test data	Ensures consistent scaling

Model Selection & Comparison

Algorithms Evaluated

Logistic Regression

- A simple **linear classification model**.
- Assumes a straight-line relationship between features and target.
- Useful as a **baseline model** to compare others against.
- Fast, interpretable, but may **underperform** when data is non-linear.

Decision Trees

- A **non-linear model** that splits data into decision rules.
- Easy to visualize and interpret.
- Captures feature interactions automatically.
- However, single trees can **overfit** and may not generalize well.

Random Forest

- An **ensemble model** of many decision trees.
- Each tree sees a random subset of data → reduces overfitting.
- Handles non-linearity, imbalanced data, and complex patterns very well.
- Typically provides **higher accuracy and stability** than single models.

Top Feature Importance

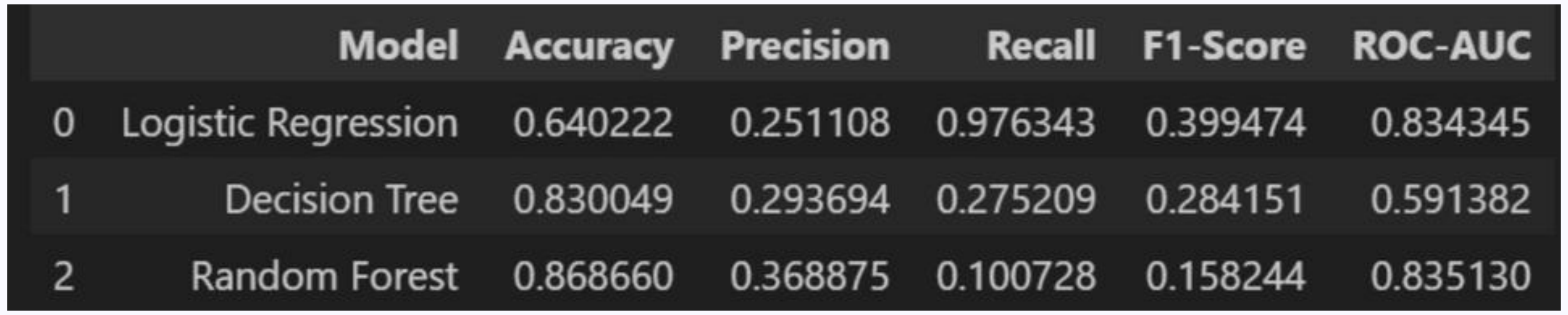
- **Vehicle Damage** → Customers with past vehicle damage show much higher interest.
- **Previously Insured** → Customers who were *not* previously insured respond the most.
- **Vehicle Age** → Older vehicles (>2 years) show higher likelihood of response.
- **Age** → Slight positive influence; responders tend to be slightly older.

Cross-validation (Model Reliability)

Applied **k-fold cross-validation** on logistic regression and random forest models.

- Ensured the model's performance is **stable, robust, and not overfitting**.
- Achieved **consistent ROC-AUC scores**, confirming good generalization.

Results & Insights



	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0	Logistic Regression	0.640222	0.251108	0.976343	0.399474	0.834345
1	Decision Tree	0.830049	0.293694	0.275209	0.284151	0.591382
2	Random Forest	0.868660	0.368875	0.100728	0.158244	0.835130

Logistic Regression gives highest recall. Random Forest gives highest accuracy but low recall.

Tuned Random Forest (Selected Model)

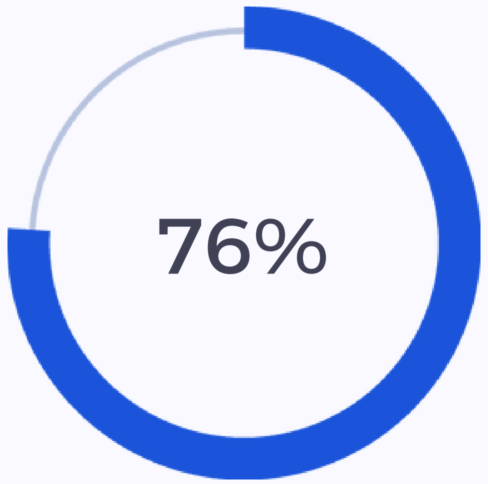
Final Hyperparameters:

- n_estimators = 300
- max_depth = 20
- min_samples_split = 5
- min_samples_leaf = 2
- class_weight = balanced
- threshold = **0.20**

Threshold Selection Summary:

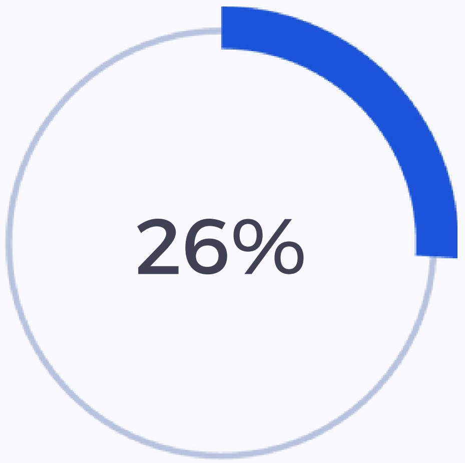
Threshold	Recall	Precision	F1
0.10	0.984	0.245	0.392
0.20	0.968	0.262	0.412
0.30	0.930	0.276	0.425

Threshold = 0.20 chosen to maximize Recall while maintaining acceptable Precision.



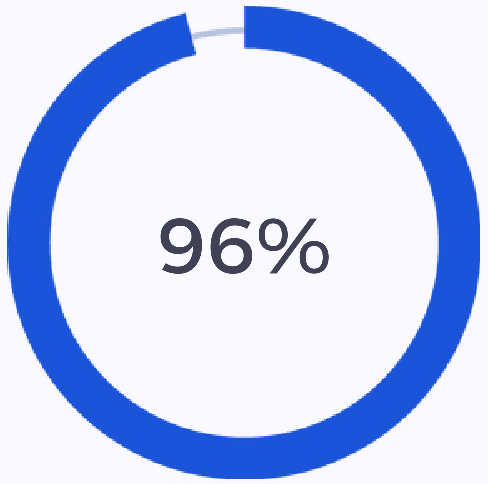
Model Accuracy

Best performing model on test data



Precision Score

Minimizing false positives



Recall Score

Capturing true positive cases

Business Value Delivered

- **Targeted Marketing:** Focus resources on high-probability customers
- **Cost Reduction:** Minimize wasted outreach to uninterested customers
- **Revenue Uplift:** Improve conversion rates through precision targeting
- **Strategic Insights:** Understand key drivers of customer interest

Limitation

Class Imbalance
Only ~12% positive responses

Low-Value Features
Driving License, Premium, Vintage add little

Missing Customer Depth
No income, claims, or communication logs

Low Precision Risk
Higher likelihood of false positives

Model Impact
Bias toward majority class expected

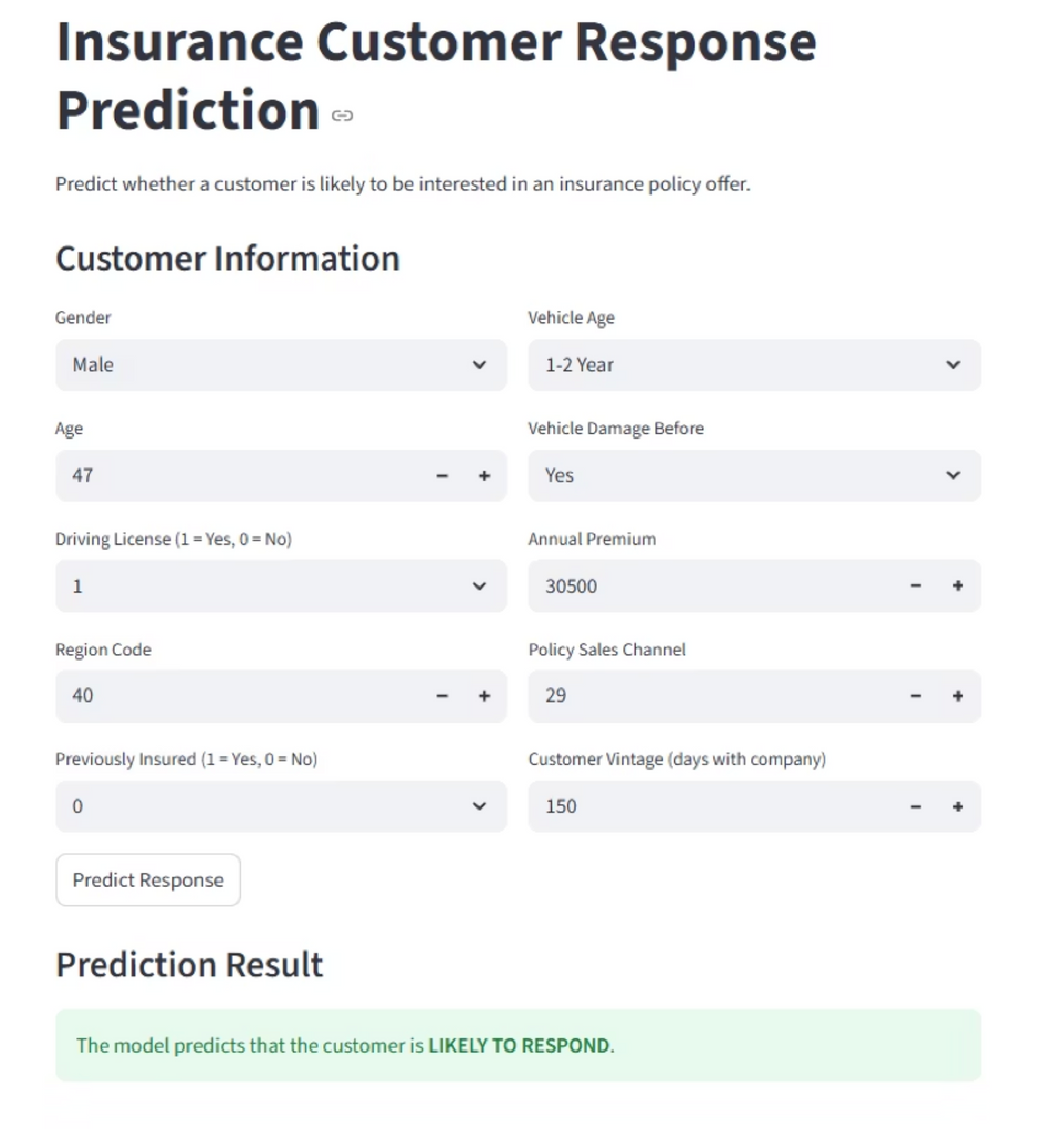
Key Findings

- Customers **not previously insured** and those with **vehicle damage** show the highest interest in purchasing insurance.
- **Vehicle Age (>2 years)** is a strong indicator of response likelihood.
- Logistic Regression achieved **high recall**, Random Forest achieved **high accuracy**, but**Tuned Random Forest (threshold = 0.20)** delivered the best balance for business needs.
- High-recall model ensures **maximum customer capture** for targeted marketing.

Future Research

- Add more meaningful features (demographics, past claims, customer interactions).
- Experiment with advanced models like **XGBoost, LightGBM, CatBoost**.
- Build a **real system** and integrate model into CRM for real-time targeting.

User Interface for Prediction



Insurance Customer Response Prediction

Predict whether a customer is likely to be interested in an insurance policy offer.

Customer Information

Gender: Male

Vehicle Age: 1-2 Year

Age: 47

Vehicle Damage Before: Yes

Driving License (1 = Yes, 0 = No): 1

Annual Premium: 30500

Region Code: 40

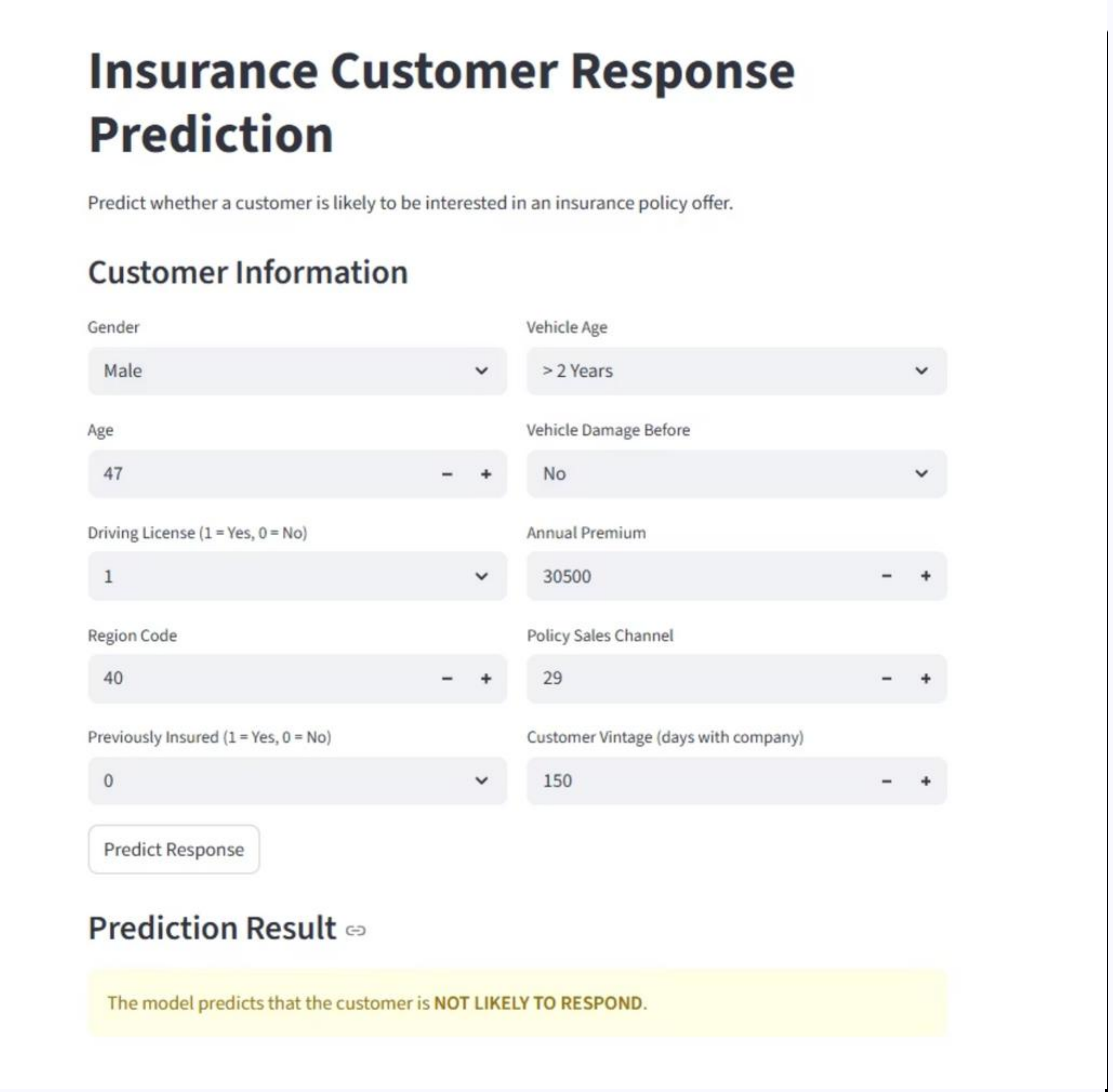
Policy Sales Channel: 29

Previously Insured (1 = Yes, 0 = No): 0

Customer Vintage (days with company): 150

Prediction Result

The model predicts that the customer is **LIKELY TO RESPOND**.



Insurance Customer Response Prediction

Predict whether a customer is likely to be interested in an insurance policy offer.

Customer Information

Gender: Male

Vehicle Age: > 2 Years

Age: 47

Vehicle Damage Before: No

Driving License (1 = Yes, 0 = No): 1

Annual Premium: 30500

Region Code: 40

Policy Sales Channel: 29

Previously Insured (1 = Yes, 0 = No): 0

Customer Vintage (days with company): 150

Prediction Result

The model predicts that the customer is **NOT LIKELY TO RESPOND**.

Thank You!