




powered by  
**bytwise.**

www.bytwiseltd.com



# ML/DL Track - Task 7

 Status	Not started
<input checked="" type="checkbox"/> Reviewed	<input type="checkbox"/>

## Report on Identifying the Best Predictor for Insurance Claims

### Introduction

In the insurance industry, predicting whether a customer will make a claim is crucial for optimizing pricing and risk assessment. This report details the process undertaken on behalf of "On the Road" car insurance to identify the single best predictor from their customer data for predicting claim occurrences using logistic regression.

### Dataset Description

The dataset provided by "On the Road" car insurance includes the following columns:

- **id**: Unique client identifier
- **age**: Client's age category (0: 16-15, 1: 26-39, 2: 40-64, 3: 65+)
- **gender**: Client's gender (0: Female, 1: Male)
- **driving\_experience**: Years of driving experience (0: 0-9, 1: 10-19, 2: 20-29, 3: 30+)

- **education:** Education level (0: No education, 1: High school, 2: University)
- **income:** Income level (0: Poverty, 1: Working class, 2: Middle class, 3: Upper class)
- **credit\_score:** Credit score (0 to 1)
- **vehicle\_ownership:** Vehicle ownership status (0: Does not own, 1: Owns)
- **vehicle\_year:** Year of vehicle registration (0: Before 2015, 1: 2015 or later)
- **married:** Marital status (0: Not married, 1: Married)
- **children:** Number of children
- **postal\_code:** Postal code
- **annual\_mileage:** Annual mileage driven
- **vehicle\_type:** Type of car (0: Sedan, 1: Sports car)
- **speeding\_violations:** Number of speeding violations
- **duis:** Number of DUIs
- **past\_accidents:** Number of past accidents
- **outcome:** Whether the client made a claim (0: No, 1: Yes)

The target variable for prediction is `outcome`.

## Methodology

To identify the best predictor, the following steps were taken:

1. **Loading the Data:** The dataset was loaded into a Pandas DataFrame.
2. **Model Training:** For each feature, a logistic regression model was trained using the formula `outcome ~ feature`.
3. **Prediction and Accuracy Calculation:** Predictions were made for each model, and accuracy was calculated as the proportion of correct predictions.
4. **Best Feature Selection:** The feature with the highest accuracy was identified as the best predictor.

## Results

The Python code used for this process is as follows:

```
import pandas as pd
import numpy as np
from statsmodels.formula.api import logit

# Load the dataset
data = pd.read_csv("car_insurance.csv")

# Initialize variables to store the best accuracy and the corresponding feature
best_accuracy = 0
best_feature = None

# Loop through each feature and fit a logistic regression model
for feature in data.columns[:-1]: # Exclude the target variable "outcome"
    formula = f"outcome ~ {feature}"
    model = logit(formula=formula, data=data).fit()
    predictions = (model.predict(data[feature]) > 0.5).astype(int)
    accuracy = np.mean(predictions == data["outcome"])
    if accuracy > best_accuracy:
        best_accuracy = accuracy
        best_feature = feature

# Create a DataFrame to store the best performing feature and its accuracy
best_feature_df = pd.DataFrame({"best_feature": [best_feature], "best_accuracy": [best_accuracy]})

# Print the best performing feature and its accuracy
print(best_feature_df)
```

The results showed that the feature `credit_score` was the best predictor with the highest accuracy. The output DataFrame containing the best feature and its accuracy is:

best_feature	best_accuracy
credit_score	0.645283

## **Conclusion**

The analysis concluded that `credit_score` is the single best predictor for whether a customer will make a claim, achieving an accuracy of approximately 64.5%. This feature can be utilized by "On the Road" car insurance to implement a simple yet effective predictive model for claim likelihood, assisting in better risk assessment and pricing strategies.

For further enhancement, it is recommended to explore additional features or combinations of features, and consider more sophisticated models to improve predictive performance.

---