Applied AI & ML Industry Projects Lab (AAM-IPL)
Transforming Learning into Industry Solutions

Week-4 : Logistic Regression

Logistic Regression Overview and Project Announcement

V Semester - ML and AIUP, Aug-Oct 2024

Session Date and Time: 5th Oct 2024 10:30 AM IST – 12:00 Noon IST

Venkateswar Reddy Melachervu
Visiting Faculty and CTO, Brillium Technologies
Department of Computer Science Engineering – AI and ML (CSM)
Email: venkat.reddy.gf@gprec.ac.in

G.Pulla Reddy Engineering College (Autonomous)
G.Pulla Reddy Nagar, Nandyal Road, Kurnool, AP 518007, India
Website: https://www.gprec.ac.in

Disclaimer and Confidentiality Notice

The content of this guest lecture, including all discussions, materials, and demonstrations, code is intended for educational purposes only and reflects the views and opinions of the speaker. While every effort has been made to ensure the accuracy and relevance of the information presented, it should not be considered as legal, financial, or professional advice.

Brillium Technologies retains unrestricted ownership of all information shared during this session. Participants must not record, reproduce, distribute, or disclose any part of the lecture or materials without prior written permission from Brillium Technologies. Unauthorized use or distribution of the content may result in legal action.

Additionally, all trademarks, service marks, and intellectual properties referenced or used in this presentation are the property of their respective owners. No ownership or rights over such third-party content are claimed or implied by the author or Brillium Technologies or by GPREC.

By attending this lecture, you agree to respect the confidentiality of the information shared and refrain from using it in any unauthorized manner. Failure to comply with these terms may result in legal action.

Thank you for your understanding and cooperation.

Lecture Outline

- Predicting Discrete or Categorical Values
- Discrete Value Predictions
- Logistic Regression Model and Logistic Function
- Prediction Probabilities
- Determining Regression Coefficients
- Model Evaluation Metrics
- RoC and Key Terms
- Vehicle Insurance Cross-Sell Prediction Project Overview
- Project Implementation Steps
- Key Implementation Advanced Concepts

Predicting Discrete or Categorical Values

- Linear Regression
  - Predicting the house value in Boston housing area
  - Predicting the value of a used e2w etc.
  - Predicting the future price of an equity stock
  - Response or inferred variable y is ***Continuous***

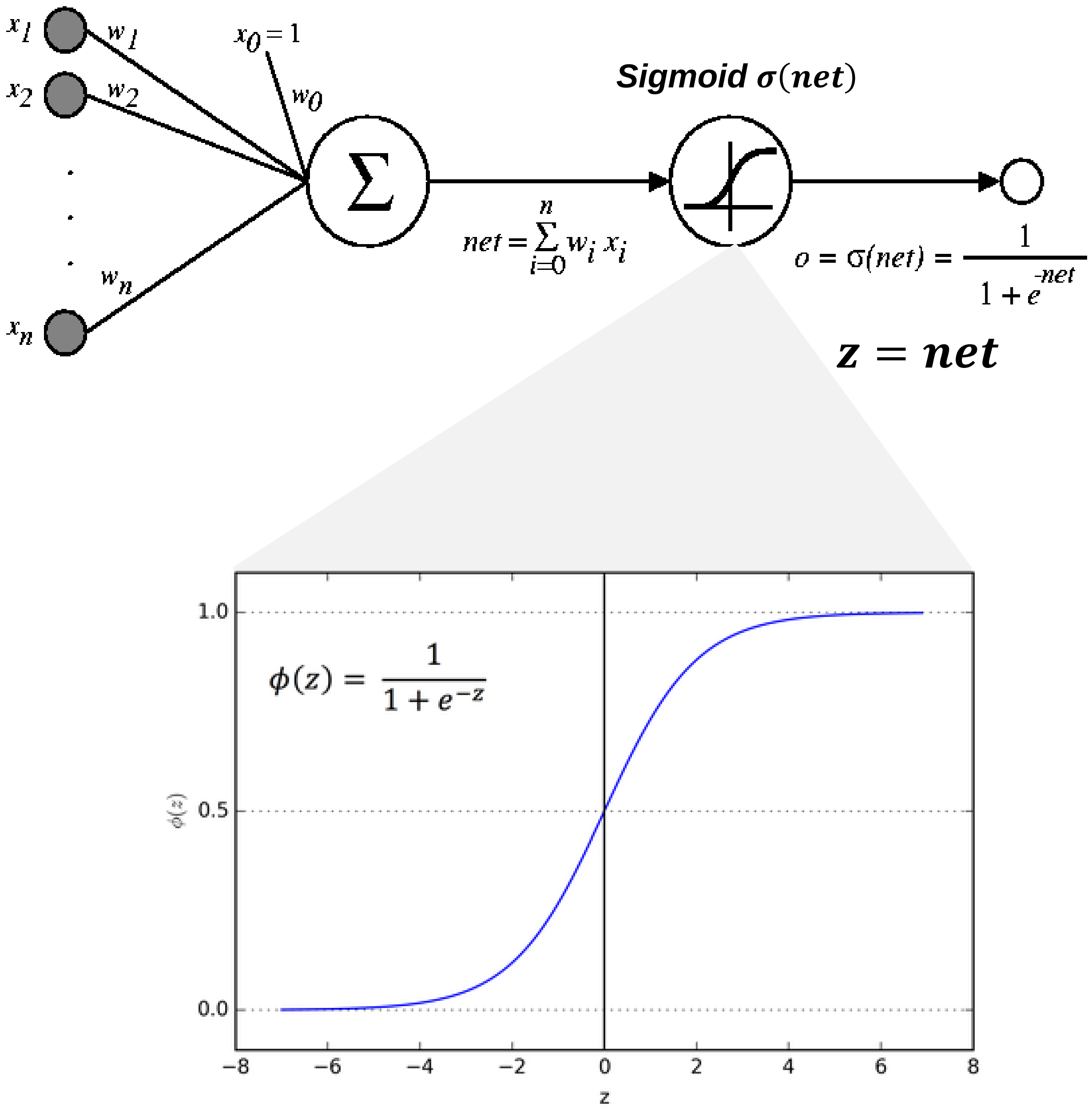
What if the response or inferred variable y is ***discrete or categorical***?

Discrete Value Predictions

- Two possible outcomes
  - Customer's interest in buying a vehicle insurance of a company - $\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}\}$
    - Response - 1 (Interested in buying – 1)
    - Response – 0 (Not Interested in buying - 0)
  - Received email is a spam or genuine - $\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}\}$
  - A credit card transaction is a fraudulent transaction or genuine transaction - $\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}\}$
  - Predicting a person has a disease (diabetes, heart-related etc.) or not based on diagnostic reports and data
  - Yes/No, True/False, Success/Failure
- Multiple possible outcomes
  - Predicting a customer's preferred product category or brand (iPhone, Google Pixel, Samsung Galaxy, OnePlus etc.)
  - Image classification into cat, dog, bird, car, bus etc.
- Is Linear Regression suitable?
  - Predicted values are continuous and unbounded
    - Not between 0 and 1
    - Not bounded
  - Hence - Not suitable
- The outcomes {0,1} are more like probabilities
- Need a model that maps real-valued numbers to a value in the range [0,1] - Binary Outcomes

Logistic Regression Model and Logistic Function

- A model that transforms the output of a linear equation **(regression)** into a probabilities **(logical or discrete or categorical)** using a special function
- The **Logistic Function** aka **Sigmoid Function** squashes linear equation output to a range between 0 and 1
  - $\boldsymbol{\sigma(z) = \frac{1}{(1+e^{-z})}}$
  - Where
    - $z = w_0x_0 + w_1x_1 + w_2x_2 + w_nx_n$ - linear combination of input features
    - $\sigma(z)$ is the probability that the output prediction belongs to value/class 1
- $\sigma(z)$:
  - 1 – The input instance belongs to class 1 (Positive class)
  - 0 – The input instance belongs to class 0 (Negative class)
  - Typical boundary condition for classification:
    - $\sigma(z) = \begin{cases} 1 & if\ \sigma(z) \geq 0.5 \\ 0 & if\ \sigma(z) < 0.5 \end{cases}$
- The Sigmoid function is:
  - Smooth
  - Continuous
  - Differentiable



Logistic Regression Model and Logistic Function

- Two types based on the outcome (dependent variable)
  - Binary Logistic Regression – two possible outcomes/classes
  - Multinomial Logistic Regression – multiple possible outcomes/classes
- Estimates the discrete probability that a given input instance belongs to a particular class – 1 or 0
- **We will focus on Binary Logistic Algorithm in this programme**

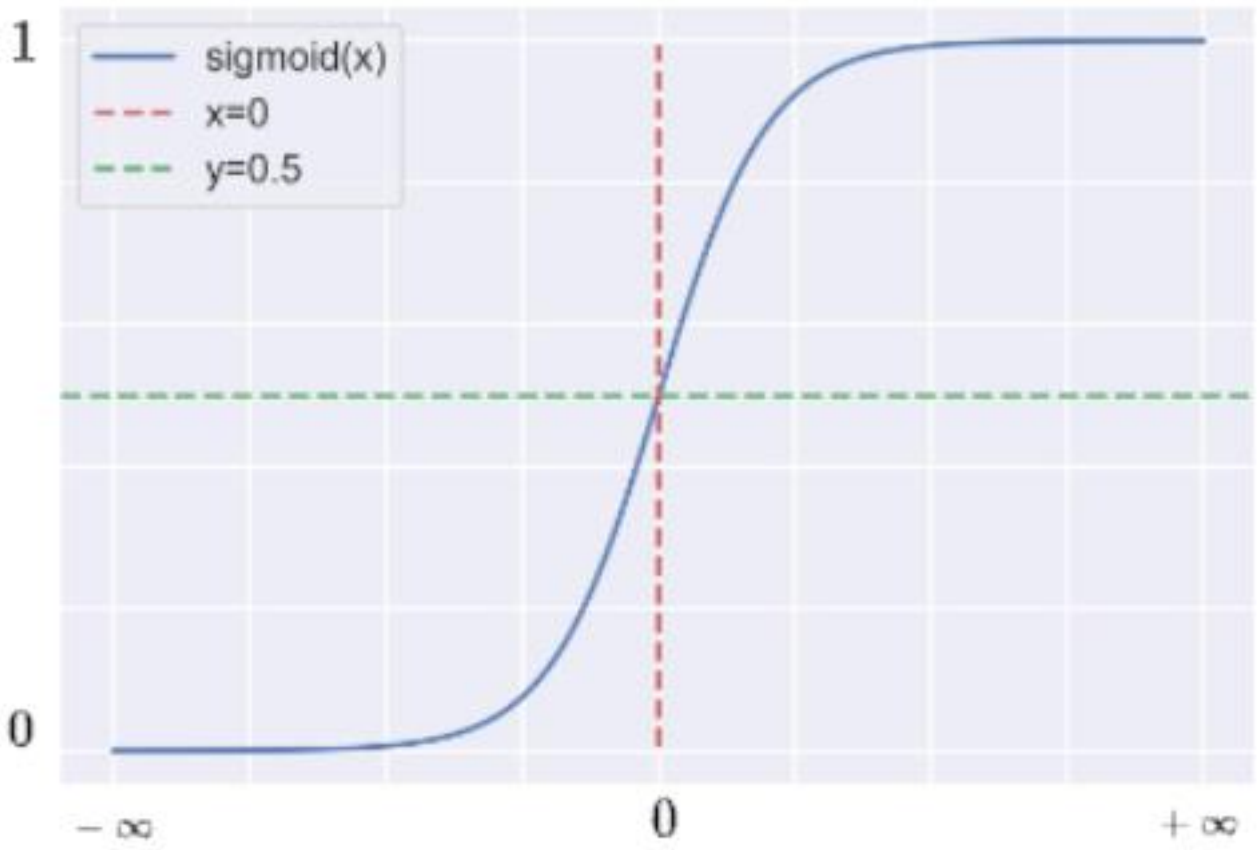
What is $\sigma(z)$ valaue for below cases

1. $\lim_{z \to -\infty} \sigma(z) = \lim_{z \to -\infty} \frac{1}{(1+e^{-z})}$ = ?
2. $\lim_{z \to \infty} \sigma(z) = \lim_{z \to \infty} \frac{1}{(1+e^{-z})}$ = ?
3. $\sigma(z) = ?$ when $z = 0$

Logistic Regression Model and Logistic Function

What is $\sigma(z)$ valaue for below cases

1. $\lim_{z\to-\infty} \sigma(z) = \lim_{z\to-\infty} \frac{1}{(1+e^{-z})} = 0$
2. $\lim_{z\to\infty} \sigma(z) = \lim_{z\to\infty} \frac{1}{(1+e^{-z})} = 1$
3. $\sigma(z) = 0.5$ when $z = 0$



Prediction Probabilities

- Input regression vector $\bar{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$

- $P(y = 1 \mid \bar{x}) = \frac{1}{\left(1+e^{-\bar{x}^T\bar{w}}\right)} = g(x) \; where \; \bar{x}^T\bar{w} = \; w_0x_0 + w_1x_1 + w_2x_2 + w_nx_n$
    - $P(y = 1 \mid \bar{x})$ – probability of response/prediction belongs to value/class 1, given input feature vector $\bar{x}$
    - y – Response/target value
    - $\bar{x}$ - Input feature vector or regression vector
    - $\bar{w}$ - Regression coefficients or weights vector

- $P(y = 0 \mid \bar{x}) = 1 - P(y = 1 \mid \bar{x}) = 1 - \frac{1}{\left(1+e^{-\bar{x}^T\bar{w}}\right)} = \frac{\boldsymbol{e^{-\bar{x}^T\bar{w}}}}{\boldsymbol{\left(1+e^{-\bar{x}^T\bar{w}}\right)}} = \mathbf{1} - \boldsymbol{g(x)}$

Determining Regression Coefficients

- Logistic regression uses ***Maximum Likelihood Estimate – MLE*** to determine the values of regression coefficients
  - Set of coefficients that maximize the likelihood of observing the actual data under the model
- The likelihood of $\left(y(k), \bar{x}(k)\right)$ $(k^{th}$ observation)$L(\bar{w}) = \left(g\left(\bar{x}(k)\right)\right)^{y(k)}\left(1 - g\left(\bar{x}(k)\right)\right)^{1-y(k)}$
- The joint likelihood of al responses and inputs $L(\bar{w}) = \prod_{k=1}^{M}\left(g\left(\bar{x}(k)\right)\right)^{y(k)}\left(1 - g\left(\bar{x}(k)\right)\right)^{1-y(k)}$
- Joint log-likelihood $L(\bar{w}) = \sum_{k=1}^{M}\left[y(k)\, ln\left(g\left(\bar{x}(k)\right)\right) + \left(1 - y(k)\right) ln\left(1 - g\left(\bar{x}(k)\right)\right)\right]$
- To maximize the log-likelihood, **gradient ascent technique** may be employed
- The regression coefficients update rule, during the training, reduces to

$\bar{w}(k+1) = \bar{w}(k) + \eta\left(y(k+1) - g\left(\bar{x}(k+1)\right)\big|_{\bar{w}=\bar{w}(k)}\right)\bar{x}(k+1)$ where $g\left(\bar{x}(k+1)\right) = \dfrac{1}{\left(1+e^{-\left(\bar{x}(k+1)^T\bar{w}(k)\right)}\right)}$

Update

Current Estimate

Learning Rate

True Response

Predicted Response

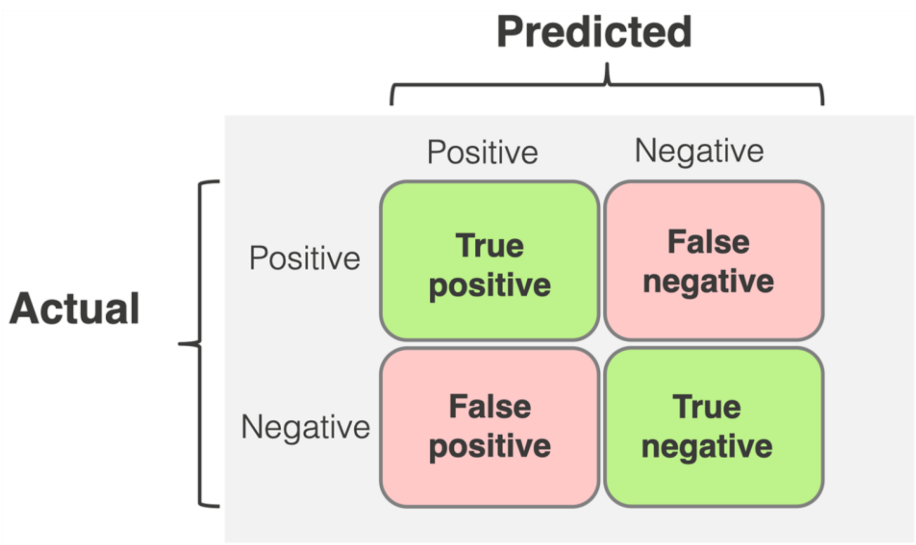
Observed Input

Prediction Error

Also Called Online Learning Algorithm

Model Evaluation Metrics

- Accuracy
    - Measures overall correctness
    - However, in imbalanced datasets, accuracy can be misleading.
- Confusion Matrix



Model Evaluation Metrics - Receiver Operating Characteristic Curve

- RoC originated from the field of **Signal Detection Theory** during WW II to evaluate radar performance
- Receiver refers to RADAR and curve was used to distinguish between **signal** (actual aircraft) and **noise** (false alarm)
- Logistic Regression is trying to do similar thing – **distinguish two classes**
- Model's performance plot across different threshold values, showing the trade-off between signal - **True Positive Rate (Recall)** and noise - **False Positive Rate**

Model Evaluation Metrics - RoC Key Terms

True Positives

- The instances where the model correctly predicts the positive class as the positive class

False Negatives

- The instances where the model incorrectly predicts the positive class as the negative class

True Negatives

- The instances where the model correctly predicts the negative class as the negative class

False Positives

- The instances where the model incorrectly predicts the negative class as the positive class

Model Evaluation Metrics – TPR, FPR, Specificity, AUC

True Positive Rate aka Recall aka Sensitivity

- Out of all actual positives, how many did the model correctly predict as positive - $\frac{TP}{TP+FN}$

False Positive Rate - FPR

- How may actual negatives are incorrectly identified as positives - $\frac{FP}{FP+TN}$

Specificity

- Out of all actual negatives, how many are correctly predict as negatives - $\frac{TN}{TN+FP}$

Area Under The Curve

- Single value summarizing the model's performance - A perfect model has an AUC of 1

Model Evaluation Metrics – F1 Score

- Harmonic mean of precision and recall
  - Harmonic Mean
    - An average calculated by diving the number of values by the sum of reciprocals of those values – useful to calculate the average of rates or ratios
    - HM = $\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
- $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
- Precision asks - "Out of the predicted positives, how many were actually positive?“
- Recall asks - "Out of the actual positives, how many did we correctly predict as positive?“
- Significance
  - If either precision or recall is very low, the harmonic mean (and thus the F1 Score) will also be low
  - This reflects that the model isn’t performing well overall if one of these metrics is very weak

Project Overview - Vehicle Insurance Cross-Sell Prediction

- Project Title
  - Cross-Sell Vehicle Insurance Prediction
- Object and Project Statement
  - A healthcare insurance company that has a sizeable existing customer for its health insurance product needs your help in building a ML to predict whether the existing health insurance customers/policyholders will also be interested in buying vehicle/automobile insurance provided by the company (2,3, and 4-wheelers) – such selling of another product to an existing customer is called cross-selling

  - Vehicle insurance is a mandatory insurance that needs to be bought by each vehicle owner and usually purchased/renewed annually.

  - Building an ML model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and increase revenue.

Project Overview - Vehicle Insurance Cross-Sell Prediction

- Dataset
  - The provided dataset - train and test CSVs – contain information about demographics (gender, age, region code type), vehicles (vehicle age, damages), healthcare policy (premium, sourcing channel) etc. The train CSV contains values for the response column whereas the values for this column are blank and need to be predicted and filled in the test CSV file.

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy_Sales_Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

Project Overview - Vehicle Insurance Cross-Sell Prediction

- Dataset
  - The provided dataset - train and test CSVs – contain information about demographics (gender, age, region code type), vehicles (vehicle age, damages), healthcare policy (premium, sourcing channel) etc.
  - **train.csv** contains values the data for training
  - **test.csv** contains the data for prediction with response column being blank
  - Training Data Set
    - Number of rows in train.csv : 3,81,109
    - Number of rows in test.csv : 1,27,037
    - Number of Features: 10
    - Response Variable (Target): Customer is interest or NOT interested (1 or 0)
  - Features
    - Gender, Age, Driving_License, Region_Code, Previously_Insured, Vehicle_Age, Vehicle_Damage
    - Annual_Premium, Policy_Sales_Channel, Vintage

Project Overview - Vehicle Insurance Cross-Sell Prediction

- Usage in Machine Learning:
  - The dataset is used to predict the binary response/interest of the customer in buying another product from the same company.
- Data Source and Published By:
  - The data is published by Anmol Kumar on Kaggle - Health Insurance Cross Sell Prediction 🏠 🏥 (kaggle.com)
- Data Download Link
  - Kaggle - Health Insurance Cross Sell Prediction 🏠 🏥 (kaggle.com)
- Submit the below
  - A new csv file "**AAM-IPL-insurance-prediction-submission.csv**" containing customer ID and predicted response for the interest in buying Vehicle Insurance, to be created for **test.csv** data points
  - PDF of the code, output, graphs etc. in a file titled "**AAM-IPL-Wk-4-LogiReg-Cross-Sell-Vehicle-Insurance-Prediction-Full-Code.pdf**"

Project Implementation Steps

- Load the train.csv and test.csv datasets using pandas
- Print the number of rows in the training and test datasets
- Check for any missing values in the datasets
- Encode categorical features
- Standardise the data and train the model
    - SMOTE technique may be used optionally for handling data imbalance
- Optional hyper parameter tuning can be employed
- Evaluate the model using Metrics – Classification report, confusion matrix etc.
- Plot RoC and Precision Curves
- Plot Feature Importance
- Prediction on Test Data and Create Submission File

SMOTE – Synthetic Minority Over-sampling Technique

- In imbalanced datasets, models tend to favor the majority class, leading to poor performance on the minority class
- SMOTE helps by creating new synthetic data points for the minority class, resulting in a more balanced dataset
- Benefits
  - Prevents Overfitting
    - By creating synthetic samples, SMOTE avoids duplicating data, reducing the risk of overfitting
  - Improves Model Performance
    - Leads to better learning from both classes, improving metrics like recall and F1-score for the minority class.

Hyperparameter Tuning

- Hyperparameters are model parameters that are set before training (e.g., C and solver in logistic regression)
- Tuning involves finding the best combination of hyperparameters to improve model performance
- GridSearchCV systematically tests different combinations of hyperparameters to find the best settings that maximize model performance, using cross-validation for reliability
- Parameters like C and solver can be employed for enhancing the model accuracy
  - C – Inverse of regularisation strength
    - Regularization is a technique used to prevent overfitting by penalizing large or complex model coefficients
    - Smaller C values (e.g., C = 0.01) imply stronger regularization, which helps prevent overfitting by penalizing large coefficients
    - Larger C values (e.g., C = 10) reduce the regularization effect, allowing the model more flexibility to fit the data.
  - Solver
    - Solver specifies the optimization algorithm used to fit the logistic regression model
    - Common solvers
      - liblinear: Suitable for small datasets and supports both L1 (Lasso) and L2 (Ridge) regularization
      - saga: Efficient for large datasets and also supports both L1 and L2 regularization. It can handle more complex problems like sparse data.

Feature Importance

- Feature Importance measures the impact each input variable (feature) has on a model's predictions. It shows which features contribute the most to predicting the target variable
- In logistic regression, feature importance is based on the coefficients of the model
  - Positive coefficients: Increase the likelihood of the target class
  - Negative coefficients: Decrease the likelihood of the target class
  - Larger absolute values: Indicate stronger influence on the prediction
- Features like Annual Premium or Vehicle Age may have higher importance in predicting whether a customer will purchase insurance (Response = 1)

Project Implementation Demo


Q & A

Interested in building a Gen AI application?

Reach out to venkat@brillium.in

THANK YOU!

AAML-IPL Brought You in Partnership with:

Brillium Technologies

Transforming Lives Through Technology

Brillium Technologies

Sector 7, HSR Layout, Bengaluru 560102, Karnataka, India
Website: www.brillium.in | Email: connect@brillium.in