\*SUMMARY of project

**Data Preprocessing:**

The code begins by reading a dataset from a CSV file named 'CarPrice\_Assignment.csv' into a pandas DataFrame named 'Cars'.

It performs one-hot encoding on categorical variables using the get\_dummies() function.

The 'price' column is sorted in descending order to observe the top 20 highest-priced cars.

Dimensionality reduction is performed by selecting specific columns/features from the dataset.

Categorical variables such as 'cylindernumber' are converted from string to integer format for numerical analysis.

**Normalization of data:**

Z-score normalization is applied to the dataset using the formula: (value - mean) / standard deviation.

Mean and standard deviation are calculated for each feature to perform normalization.

**Linear Regression:**

Data is split into features (X) and the target variable (y).

Linear regression is performed using the formula: y = mx + c, where 'm' is the slope and 'c' is the intercept.

The slope and intercept of the regression line are calculated.

Predictions are made on new data using the calculated slope and intercept.

Finally, the least fit curve (regression line) is plotted using

matplotlib.

GITHUB LINK—-

<https://github.com/rampofin/mrm.git>

Review questions-

1.Difference between loss and cost function.

Loss function—

* Measures error between predicted and actual values for individual data points.
* Used during model training to adjust parameters to minimize error.
* A convex loss function has only one global minimum and no local minima, making it easier to solve with a simpler optimization algorithm. However, a non-convex loss function has both local and global minima and requires an advanced optimization algorithm to find the global minimum.
* Examples include Mean Squared Error (MSE), Cross-Entropy Loss, Hinge Loss.

**Cost function—**

* Represents average of loss function over entire training dataset.
* Evaluates overall model performance and guides optimization process.
* Often computed as the average of individual loss functions across all training examples

2.Vanishing gradient descent

Vanishing gradient problem is a phenomenon during which gradients that are used to update the network become very small or "vanish" as they propagates backward from the output layers to the initial layers.

The consequences of vanishing gradient problem include slow convergence, network getting stuck in low minima, and impaired learning of deep representations.

3. Normalisation

* Data normalization is a crucial preprocessing technique used for forecasting and prediction models to enhance accuracy. It involves transforming the current data range into a new, standardized range.
* Normalization is particularly important for aligning different prediction and forecasting techniques effectively. It improves the consistency and comparability of various predictive models by standardizing the range of independent variables or features within a dataset.
* Normalization reshapes numerical columns to adhere to a standard scale, especially useful for datasets with varying units or magnitudes across features. The primary goal is to find a common scale for the data while preserving intrinsic variations in value ranges.
* Common normalization methods include:
  + Rescaling features to a standard range, typically between 0 and 1.
  + Adjusting features to have a mean of 0 and a standard deviation of 1.

4.Assumptions taken during Logistic and Linear Regression

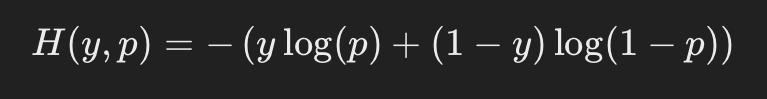
**\*Linear Regression Assumptions**:

* **Linearity**: The relationship between the independent variables (features) and the dependent variable (target) is linear. This means that the change in the dependent variable is proportional to the change in the independent variable.
* **Independence**: The observations are independent of each other. In other words, the value of one observation does not depend on the value of another observation.
* **Homoscedasticity**: The variance of the errors is constant across all levels of the independent variables. This ensures that the spread of the residuals is consistent throughout the range of the predictor variables.
* **Normality of Residuals**: The residuals (the differences between observed and predicted values) are normally
* distributed. This assumption is important for hypothesis testing and confidence intervals.
* **No multicollinearity**: The independent variables are not highly correlated with each other. High multicollinearity can lead to unstable estimates of the coefficients.

**\*Logistic Regression Assumptions**:

* **Binary Outcome**: Logistic regression is typically used for binary classification problems, where the dependent variable is categorical with two levels (e.g., 0/1, yes/no).
* **Independence**: The observations are independent of each other, similar to linear regression.
* **Linearity of Log Odds**: The relationship between the independent variables and the log odds of the outcome is linear. This assumes that the natural logarithm of the odds of the outcome is a linear combination of the independent variables.
* **No multicollinearity**: Similar to linear regression, logistic regression assumes that the independent variables are not highly correlated with each other.
* **Large Sample Size**: Logistic regression performs better with a larger sample size. While there's no strict cutoff, having a sufficiently large sample size helps ensure stable estimates of the model parameters.

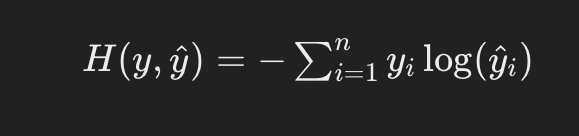
5.Cross Entropy formula

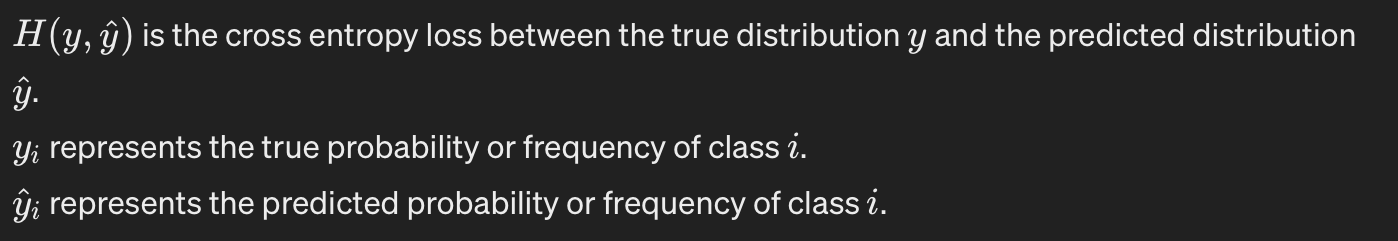
The formula for cross entropy H(y,p) is-

*y* as the actual binary label (0 or 1).

*p* as the predicted probability of the positive class (class 1).

1−*p* as the predicted probability of the negative class (class 0).

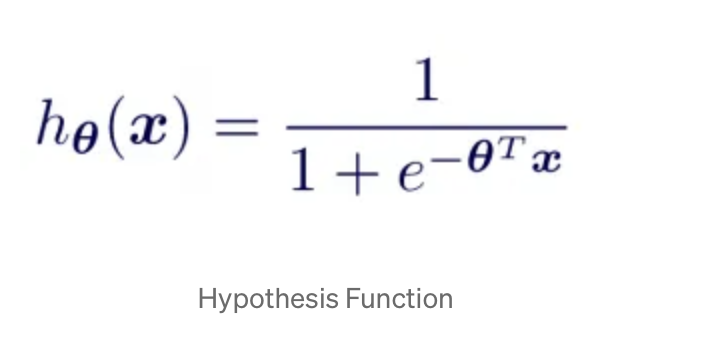
Cross entropy for multi-class classification



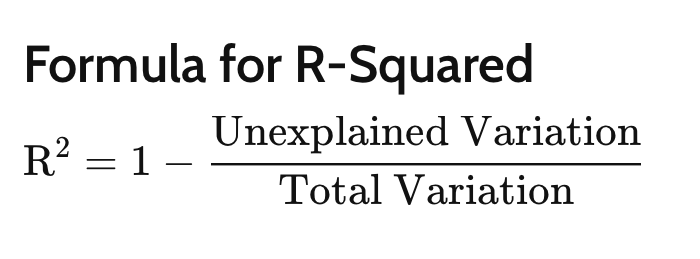
The sum is taken over all the classes(categories) in the

distribution

6.derivative of log loss function

In order to preserve the convex nature for the loss function, a [log loss](https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropy-aka-log-loss-the-cost-function-used-in-logistic-regression/?utm_source=Backlink&utm_medium=SEO) error function has been designed for logistic regression.

7.R-squared value

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable in a regression model.

8.metrics for logistic regression

**\*Accuracy:** The proportion of correctly classified instances out of the total instances. It is a straightforward measure but can be misleading when classes are imbalanced.

\***F1 Score:** The harmonic mean of precision and recall, balancing both precision and recall. It's particularly useful when classes are imbalanced.

\***ROC AUC Score (Receiver Operating Characteristic Area**

**Under the Curve)**: It measures the area under the ROC curve, which plots the true positive rate vs the false positive rate. It evaluates the model's ability to distinguish between positive and negative instances across various threshold settings.

\***Confusion Matrix**: A tabulation of actual vs predicted class

labels. It provides insights into the performance of the classifier, showing true positives, true negatives, false positives, and false negatives.

9.Batch and Stochastic Gradient Descent