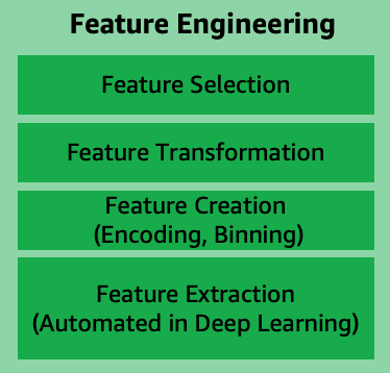
Load Data

Data cleaning – remove unwanted columns

* Handle missing values
* Remove duplicates

Feature understanding and explore



Feature relationships

Split data

Model training

Prediction

Evaluation

Model load

OpenCV

Reinforcement

LLM

LSTM

CCN

Image processing

NLP

Tensorflow

keras

One way to select features is by using domain knowledge.

Another approach involves feature extraction and selection techniques such as correlation analysis, principal components analysis (PCA), or recursive feature elimination (RFE). These techniques help you identify the most relevant features of the model while ignoring irrelevant or redundant ones.

* You need to explore the data. You must look at the distribution of the features and see if there are any outliers or missing values. You may need to clean the data or remove features that are not informative.
* You need to use feature selection techniques. Several [feature selection techniques](https://www.projectpro.io/article/feature-selection-methods-in-machine-learning/562), such as filter, wrapper, and embedded methods, are available. Each technique has strengths and weaknesses, so you must choose the most appropriate for your problem.
* You must evaluate the results. Once you select a set of features, you must assess the results. How well does the model perform with the selected features? Could you remove any features without significantly impacting the model's performance?

1. Imputation – handling missing data

deleting records missing specific values

* Categorical Imputation: Missing categorical variables are generally replaced by the most commonly occurring value in other records
* Numerical Imputation: Missing numerical values are generally replaced by the mean of the corresponding value in other records

1. Discretization

taking a set of data values and grouping sets of them together logically into bins (or buckets).

Binning can apply to numerical values as well as to categorical data values. This could help prevent data from overfitting but comes at the cost of loss of granularity of data. The grouping of data can be done as follows:

* Grouping of equal intervals
* Grouping based on equal frequencies (of observations in the bin)
* Grouping based on decision tree sorting (to establish a relationship with target)

1. Categorical Encoding

encode categorical features into numerical values

One hot encoding(OHE) - categorical values are converted into simple numerical 1’s and 0’s without losing information. As with other techniques, OHE has disadvantages and must be used sparingly. It could dramatically increase the number of features and result in highly correlated features.

Besides OHE there are other methods of categorical encodings, such as-

* Count and Frequency encoding- captures each label's representation,
* Mean encoding -establishes the relationship with the target, and
* Ordinal encoding- the number assigned to each unique label.

1. Feature splitting
2. Handling Outliers
3. Removal: The records containing outliers are removed from the distribution. However, the presence of outliers over multiple variables could result in losing out on a large portion of the datasheet with this method.
4. Replacing values: The outliers could alternatively bed treated as missing values and replaced by using appropriate imputation.
5. Capping: Capping the maximum and minimum values and replacing them with an arbitrary value or a value from a variable distribution.

6. Variable Transformation

help with normalizing skewed data.

Logarithmic transformations operate to compress the larger numbers and relatively expand the smaller numbers. This, in turn, results in less skewed values, especially in the case of heavy-tailed distributions.

Other variable transformations used include Square root and Box-Cox transformations, which generalize the former two.

7. Scaling (feature normalization)

* **Min-Max Scaling-** This process involves rescaling all values in a feature from 0 to 1. In other words, the minimum value in the original range will take 0, the maximum value will take 1, and the rest of the values between the two extremes will be appropriately scaled.
* **Standardization/Variance Scaling-** All the data points are subtracted by their mean, and the result is divided by the distribution's variance to arrive at a distribution with a 0 mean and variance of 1.

It is necessary to be cautious when scaling sparse data using the above two techniques as it could result in additional computational load.

1. **Feature Creation in Machine Learning**

Feature creation involves deriving new features from existing ones. This can be done by simple mathematical operations such as aggregations to obtain the mean, median, mode, sum, or difference and even product of two values. Although derived directly from the given input data, these features can impact the performance when carefully chosen to relate to the target

* **Benchmark**: A benchmark model is the most user-friendly, dependable, transparent and interpretable model against which you can measure your own. It’s a good idea to run test data sets to see if your new machine learning model outperforms a recognized benchmark. These benchmarks are often used as measures for comparing the performance between different machine learning models like neural networks and support vector machines, linear and non-linear classifiers or different approaches like bagging and boosting.

**Feature Engineering Tools to Know**

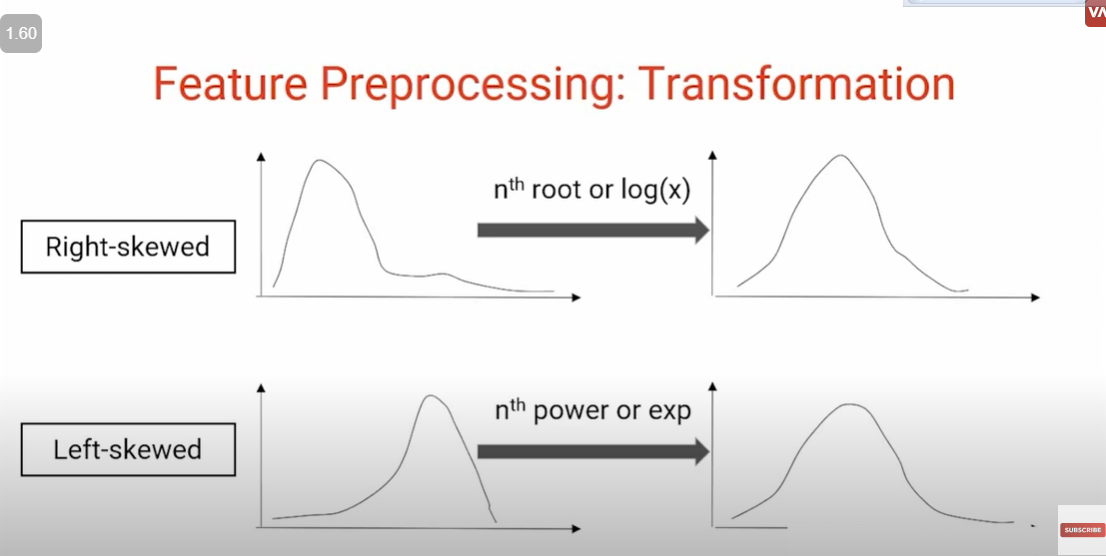
**FeatureTools**

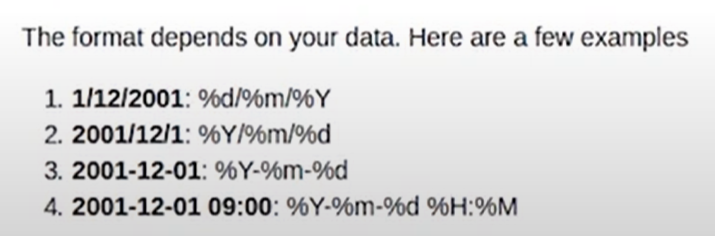
**AutoFeat -** perform linear prediction models with automated feature engineering and selection.

**TsFresh -** for time series classification and regression.

**OneBM**

**ExploreKit**

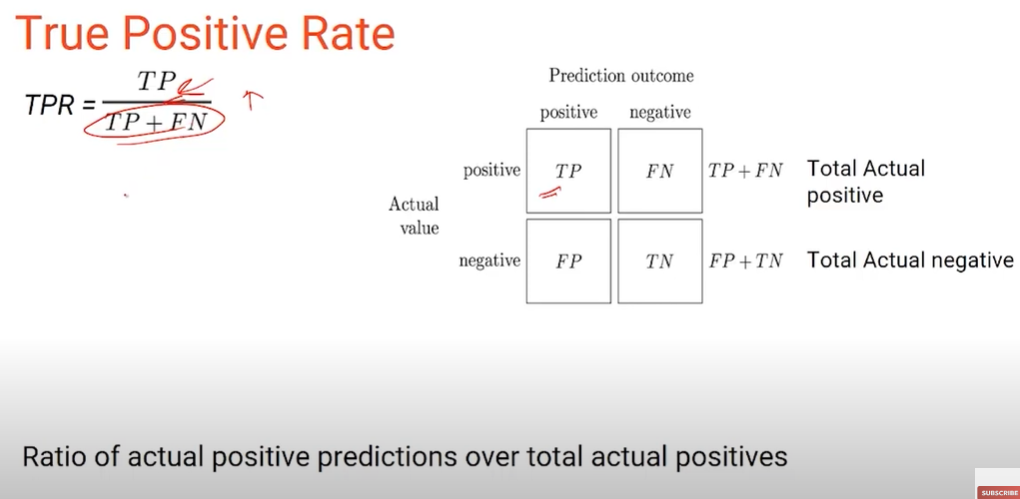
****

****

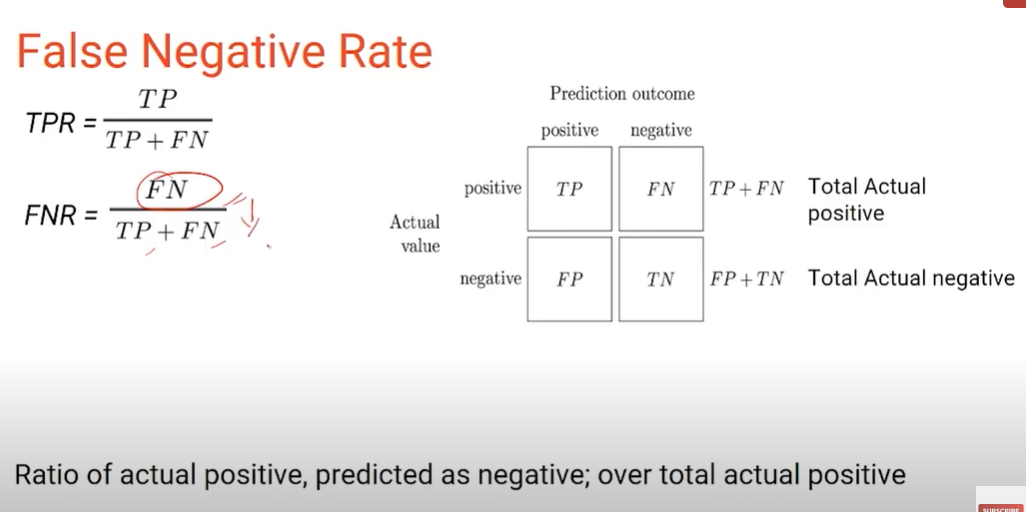
Evaluation Matrix

Confusion matrix

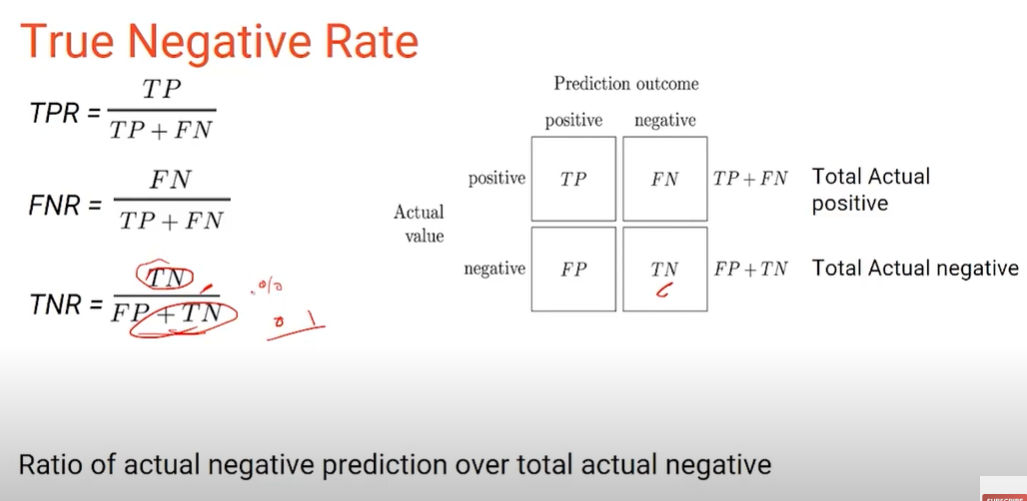
Accuracy



Higher the value better the model



Lower the value better the model



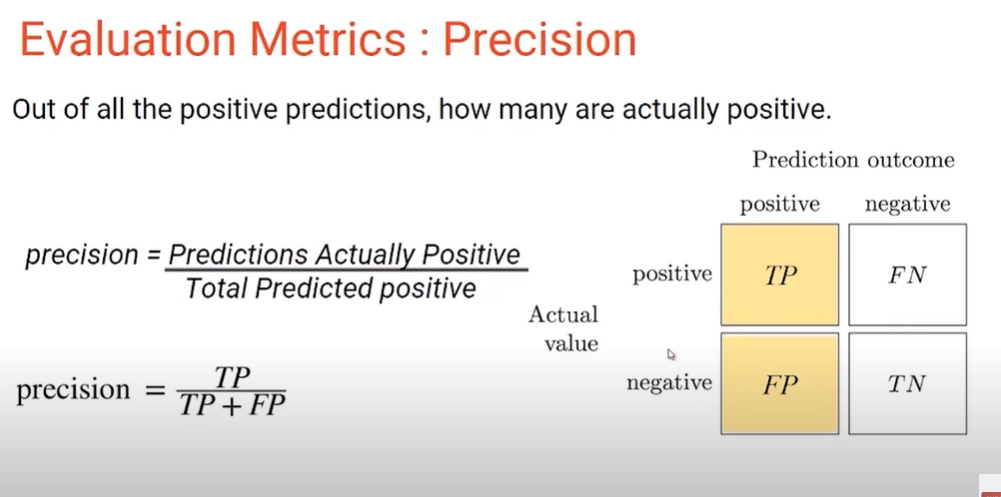
Higher the better the model

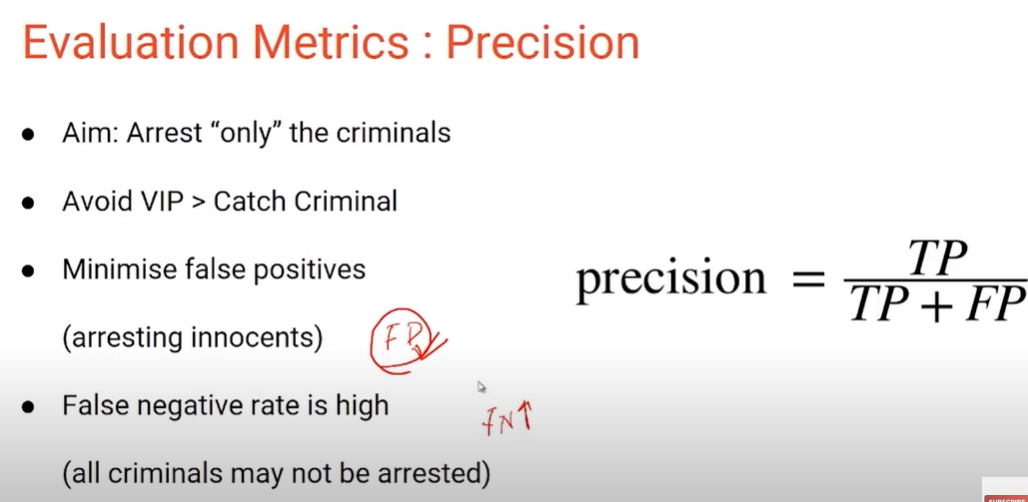
A screenshot of a computer screen

Description automatically generated

Lower the value better the model

Precision and recall

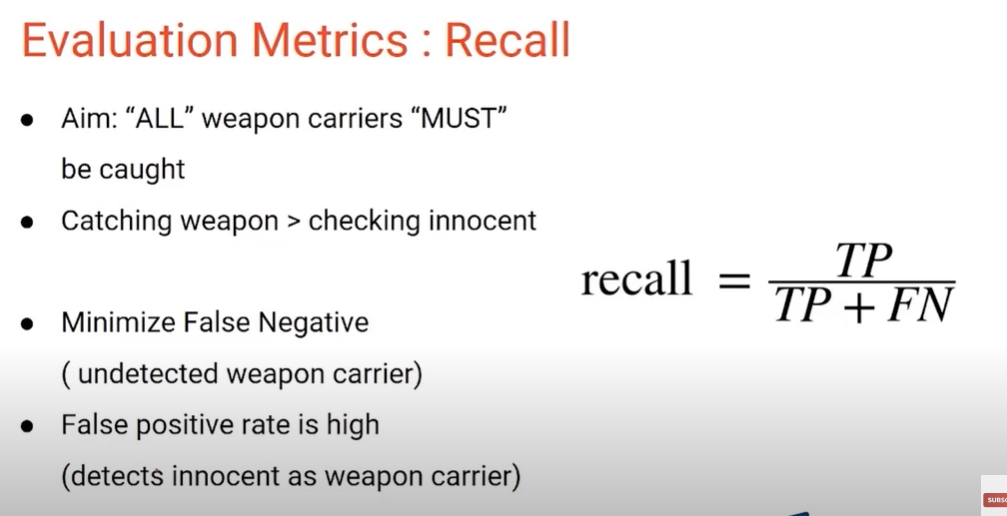




Avoid false positive

A diagram of positive results

Description automatically generated



Use recall when we cannot have false positive or avoid false negative

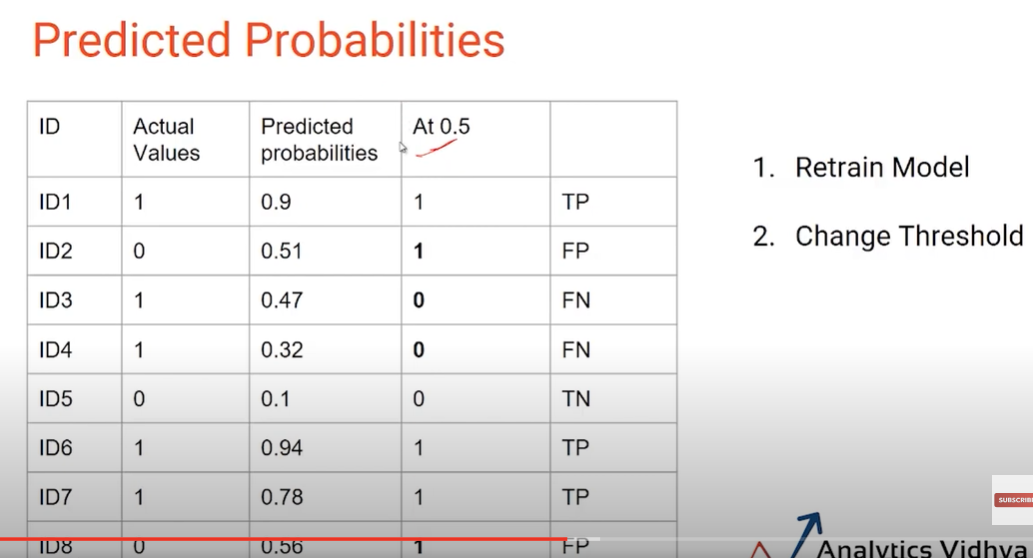
A graph on a white background

Description automatically generated

Thresholding

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a video

Description automatically generated

AUC-ROC (Area under the curve – Receiver Operating Characteristic)

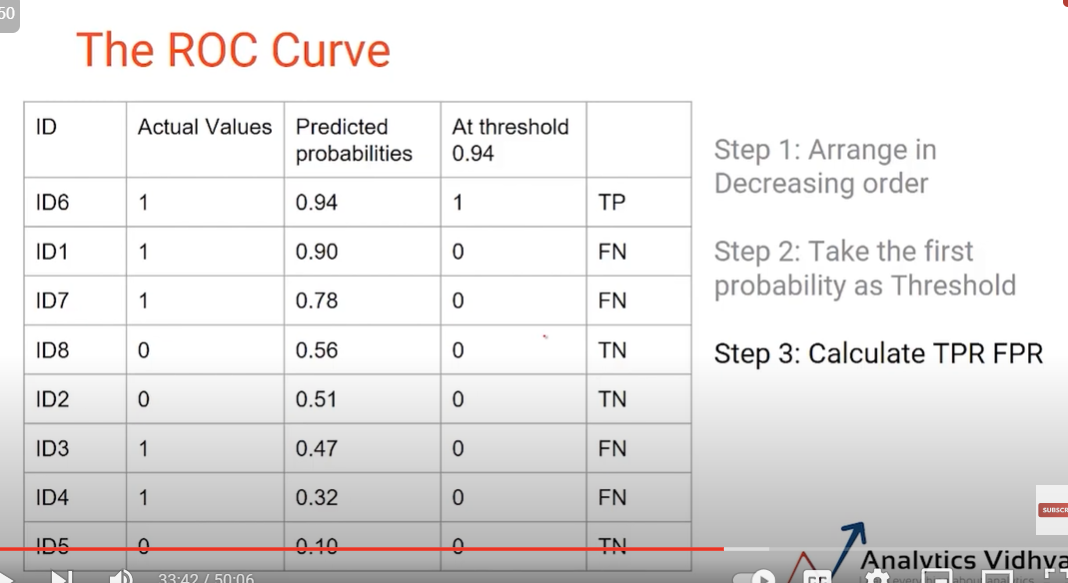
A screenshot of a computer screen

Description automatically generated

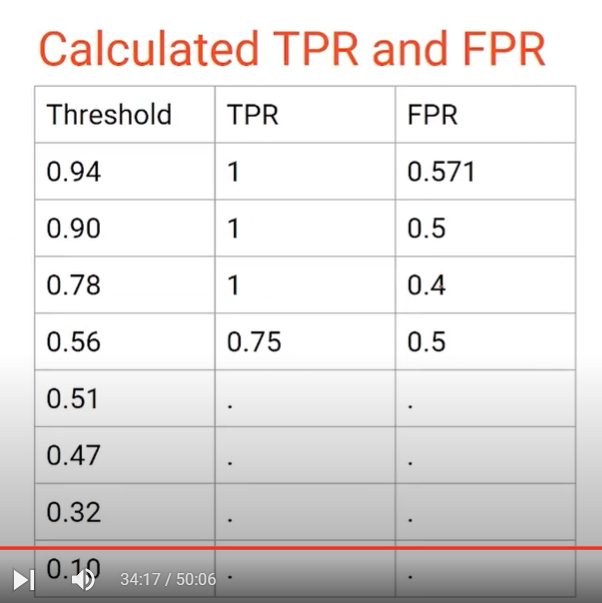
A screen shot of a graph

Description automatically generated

More the area under the curve better the model



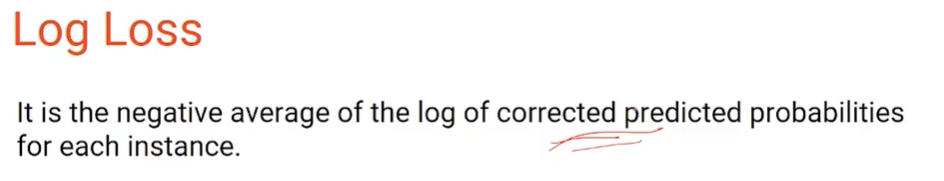
Repeat the above steps for other predicted probabilities



Log Loss

A white background with black text

Description automatically generated



A screenshot of a computer screen

Description automatically generated

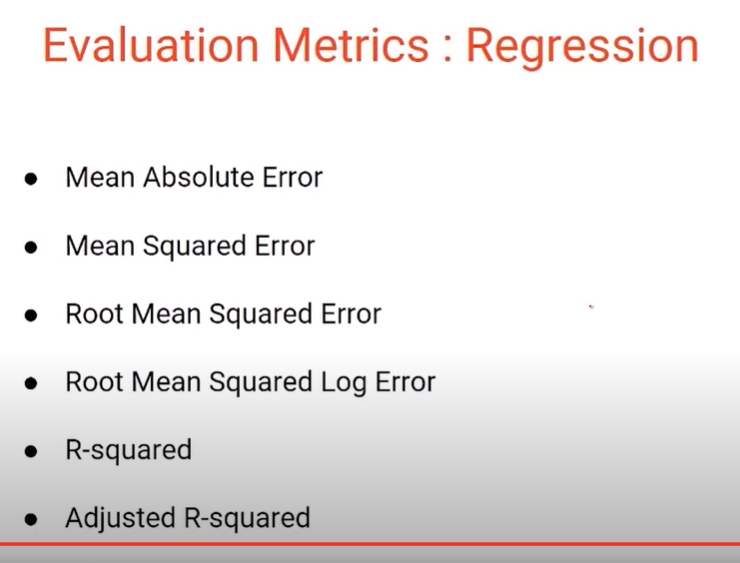
A screenshot of a video

Description automatically generated

Lower log-loss indicates better model performance.

MAE and MSE

(for regression models)(the above metrics can be used for binary models)



MAE and MSE , RMSE and RMSLE lower the better