DATA PREPARATION

BUSINESS KNOWLEDGE

[Research]

Most important part of creating a model is to have sound business knowledge of the problem.

1. Primary Research - Gather information on your own
    a. Discuss with the customers and stakeholders who are directly affected by the product. Also discuss with the other departments like marketing department and R&D department.
    b. Perform a dry run. Try to sell or buy your own product all by yourself to understand a customer perspective.
2. Secondary Research
    a. Use the information produced by others.
    b. Read reports and studies by other businesses of the same industry.
    c. Read the previous works and findings related to your problem.

DATA EXPLORATION

[Data Integration]

After doing a thorough research on the business problem and gathering ample business knowledge, next step will be to collect data and create a dataset.

1. Internal Data
    - Data collected by your organisation.
    - Examples are usage, sales and promotion data etc.
2. External Data
    - Data acquired from external data resources.
    - Examples are census data, external vendor data and scrape data etc.

[Data dictionary is used to provide detailed information about the contents of a dataset or database, such as the names of features, their data types or formats, and text descriptions. A data dictionary provides a concise guide to understanding and using the data.]

EXPLORATORY DATA ANALYSIS [EDA]

1. EDA is used to perform a deep analysis of the numerical features in our dataset.
2. The task is to find trends and use graphical plots to get insights.
3. EDA gives us a basic understanding of the data to know which steps are to be taken for data preparation.

OUTLIER TREATMENT

1. Outlier is an observation that appears far away and diverges the overall pattern of the data.
2. Outliers can be observed using percentile and graphical plots like scatter plots.
3. Reasons for their presence —
    1. Data entry
    2. Measurement error
4. Outliers can be treated using 3 commonly used methods —
    1. Capping and Flooring
        1. Observing outliers using graphical plot.
        2. All values above and below 99percentile and 1percentile are outliers.
        3. All values above 99percentile = n*99percentile, Example, 2*99percentile
        4. All values below 1percentile = 0.n*1percentile, Example, 0.2*1percentile
    2. Exponential Smoothing
        1. Observing outliers using graphical plot.
        2. Extrapolate a curve between 95 to 99 percentile
        3. Make all value above 99 percentile to fall on this curve
        4. Same for outliers below 1 percentile.
    3. Sigma Approach

- Sometimes, we might plot a graph to check for outliers but see that there is a polynomial relationship between feature and output. Example, we have three data points 1, 10 and 100. According to our normal logic, we might think that 100 is an outlier but it is not.

- This problem can be solved using multiple mathematical functions like log, exponential or root functions. For above example, if we use log to the base 10, new values will be 0, 1 and 2.

IMPUTATION

1. Dataset can have missing values that were not recorded or got corrupted.
2. This problem can be solved by replacing the missing values with harmless values -
    1. Mean or Median for numerical features
    2. Mode for categorical features
3. Another solution can be to remove the complete row which is only recommended when we have large dataset.
4. We can also impute the values with relevant values. For example - rainfall in different cities of the same state is almost similar. So we can enter the data of one city to another within the same state.
5. There are many machine learning models to impute missing values based on other features and there dependancy on the output.

BIVARIATE ANALYSIS

1. Bivariate analysis is used to study relationship between 2 features. This data can help us to see if there is any significant relationship between 2 features and use it to build an appropriate model.
2. This can be done by plotting scatter plots, correlation tables etc.
3. Scatter plots will tell if there is linear or non-linear relation between 2 features.
4. If 2 features are highly correlated to each other, then we can remove one of them with less influence on the output. Also if an independent variable has no impact on dependant variable, we can discard that variable. This can be done using correlation table.

DATA TRANSFORMATION

1. Transforming existing data into more useful form of data to extract more information.
2. We can do this by using multiple methods —
    a. Using mean or median of multiple variables showing similar type of data.
    b. Transform variable by taking exponential, x^n, root, logarithmic etc.
    c. Create ratio variables. Example: Quality of education will depend better on number of teachers per thousand students rather than number of teachers only.
3. It is also necessary to drop columns that have no relevance like
    1. Columns with same value throughout.
    2. Columns with low fill rate, when most of the rows in a column are empty.
    3. Columns with no business sense. Example, scatter plot of a feature with respect to output shows no significant relationship.

ENCODING CATEGORICAL DATA

1. Most of the machine learning models do not process categorical data. It is only possible for them to process numerical data. So it is essential to convert the categorical columns in dataset to numerical columns. This method is known as encoding.
2. There are many types of encoding methods like one-hot encoding, label encoding, ordinal encoding and binary encoding.
3. One-Hot Encoding —
    1. Imagine you have a fruit category [apple, banana, orange]
    2. One-hot creates a new binary feature for each category.
    3. So, three features (apple, banana, orange) become three binary columns.
    4. Each row shows which fruit is present (1) and the others are absent (0).
    5. Great for situations where order does not matter. Example — types of fruit
    6. Can cause curse of dimensionality.

4. Label Encoding —
    1. Assigns a unique number to each category [apple = 1, banana = 2, orange = 3]
    2. Simpler than one-hot encoding but numbers might be interpreted as order (1 < 2 < 3), which may not be true for fruits.

5. Ordinal Encoding —
    1. Similar to label encoding, but only use it if there is a natural order.

2. Example — T-shirt size [S = 1, M = 2, L = 3] where there is a clear order of size.

6. Binary Encoding [Less Common] —
    1. A mix of label and one-hot encoding.
    2. Convert label encoding to binary. Example, apple = 001, banana = 010.
    3. Less common than one-hot encoding, but can be useful for memory efficiency with many categories.

CORRELATION TABLE

1. A correlation table is a table showing correlation coefficients between all numerical features.
2. A positive value shows direct correlation and negative value shows inverse correlation.
3. Strength varies from 0 to 1, 1 being high correlation.
4. High correlation between 2 features means that there is a strong linear relationship which can lead to multi-collinearity.
5. Multi-collinearity makes the model unable to interpret impact of each individual feature on the output.
6. High correlation gives us an idea of presence of multi-collinearity but a more solid statistic to check for the presence of multi-collinearity is variance inflation factor.
7. Variance Inflation Factor
    1. VIF is used to detect the presence of multi-collinearity. It measures how much the variance of the estimated regression coefficients are inflated.
    2. It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

$$VIF = 1 / (1 - R^2)$$

    3. High R^2 results in high VIF and high VIF indicates multi-collinearity [typically >5]
8. Probable Solutions
    1. Either drop one or combine both variables based on business knowledge.
    2. Feature selection or dimensionality reduction methods.
    3. Cross-validation by checking accuracy of model with both correlated variables together and separately.

STANDARDISATION / NORMALISATION

1. Example, we have a house price dataset with 2 features, number of rooms and rainfall. Number of rooms column has values like 2, 3, 4 and rainfall (in cms) has values like 100,120,130. Now the machine learning model only understands numbers and rainfall values will be much higher than rooms values for the model. Hence, the model will interpret that rainfall as a feature has much more significance than rooms. But practically, the number of rooms will obviously have more relevance than rainfall.
2. Standardisation solves this problem. It makes sure that no single feature becomes more relevant just because of a larger scale.
3. Standardisation is a method of rescaling all the features in the dataset to a similar scale or range.
4. Various methods
    1. Min-Max scaling — rescales features between 0 and 1.
    2. Z-score normalisation — features shifted to have mean 0 and scaled by standard deviation.
    3. Robust scaling — rescales features based on percentiles.
    4. Log transformation — rescales features by logarithmic transformation.
5. Application
    1. Min-Max scaling is a good general-purpose option, but it can be sensitive to outliers.
    2. Z-score normalisation is preferred when features follow a normal distribution.
    3. Log transformation is useful for skewed data or features with a large positive range.
    4. Robust scaling is robust to outliers but might not be suitable for all algorithms.

RESAMPLING

1. Most of the machine learning models when trained on an imbalanced dataset tend to become biased towards the majority class ignoring the minority samples. This problem is solved using resampling.

2. Resampling methods are methods used to solve the problem of imbalanced datasets by creating new samples or reducing existing samples to equalise the number of samples for all classes in a dataset.
3. Resampling can be done in 2 ways — undersampling and oversampling.
4. Undersampling
    1. A method where samples of the majority class are dropped from the dataset and reduced to the number of samples present for minority class.
    2. Undersampling is only preferable when the dataset is big enough that losing some data or information is reasonable.
5. Oversampling
    1. A method where new synthetic samples of minority class are created from the already existing samples and expanded to the number of samples present for majority class.
    2. Oversampling is a preferable method when the dataset is small and losing information will make the model unable to understand the underlying trend.
6. Various methods
    1. Random undersampling
        1. Randomly drops samples of majority class until a more balanced dataset is achieved.
        2. This method is simple and fast but often leads to loss of valuable information of the majority class.
    2. Random oversampling
        1. Randomly duplicates samples of the minority class until a more balanced dataset is achieved.
        2. This method is simple and fast but often leads to an overfit model because it only duplicates the already existing samples.
    3. SMOTE [synthetic minority oversampling technique]
        1. Creates new samples called synthetic samples of the minority class using KNN.
        2. Randomly chooses a sample from minority class, finds k nearest neighbours, randomly chooses a neighbour and creates a synthetic sample along the line segment joining randomly chosen sample and neighbour.
        3. SMOTE is a good option for oversampling when compared to Random Oversampling but can often create synthetic samples that are not possible to exist for real-world data.
    4. Cluster based oversampling
        1. This method also uses SMOTE to create synthetic samples but before that, this method groups samples of minority class into clusters. This solves the problem of just using SMOTE where all minority samples are treated equal.
        2. This method can create more relevant synthetic samples than just SMOTE. On the other hand, appropriate clustering method is to be chosen to create clusters and can be computationally expensive.

TRAIN-TEST SPLIT

1. Testing the data is one of the many essential tasks in model creation. If the complete dataset is used to train the model and have no data left to test it, it is impossible to validate the performance and optimise the model.
2. There is a need to test the trained model on a previously unseen data because of a high possibility of getting a low training error and high testing error at the same time.

Example - We want to find a medical problem in a person who has not yet been diagnosed and not in a person who has already been diagnosed.

1. Validation Set Approach
    1. Dataset is randomly split into two parts — a training set and testing set.
    2. Model is trained on training set and evaluated on testing set.
    3. A commonly used split for train:test is 80:20
    4. Limitation — a part of data will not be used to train.

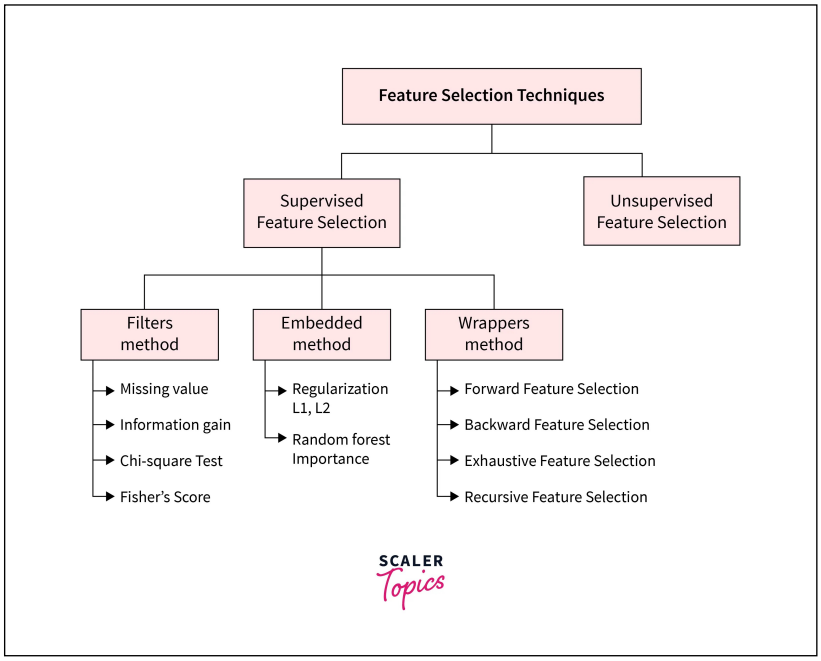
2. K-Fold Cross Validation
    1. Dataset is divided into k number of equal folds.
    2. Each fold is kept aside for testing once and the rest of them are used for training.
    3. We can calculate means and standard deviations for k testing results and understand how much our evaluation can vary when another sample is used.

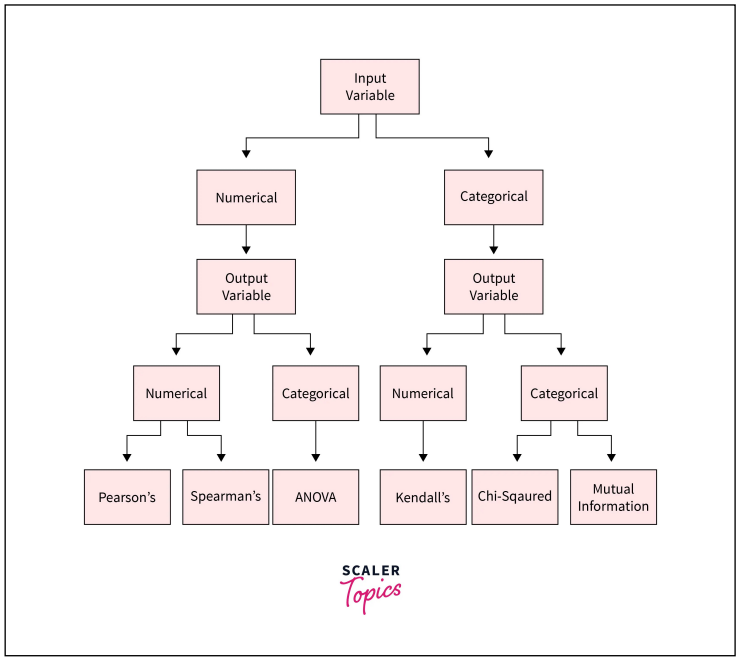
3. Leave-One-Out Cross Validation
    1. Special case of k-fold cross validation where k is equal to number of samples.
    2. We keep the first observation as test case and use rest of the observations for training purpose. Then we keep the second observation as a test case and use the rest of them for training purpose. This goes on for k number of cycles.
    3. High computational cost

4. Stratified K-Fold Cross Validation
    1. This method is another extension of k-fold cross validation and is good for imbalanced datasets.
    2. Same as k-fold cross validation but this method makes sure that each fold created has equal proportion of classes. This prevents the model from being biased towards majority class.

FEATURE SELECTION

1. https://www.scaler.com/topics/machine-learning/feature-selection-in-machine-learning/





WRAPPER METHODS

1. Wrapper methods are methods of selecting or reducing to a good number of features to enhance the model performance while training.
2. Features that have not much contribution or impact on the output act as noise while training a model which can lead to an overfit model.
3. Best Subset Selection Method | Exhaustive Feature Selection
    1. Assuming we have p features in the dataset
    2. In the first step, the model will train with no feature and simply give the mean value and save it as M0
    3. In the second step, the model with train with 1 feature at a time and select the best result [lowest RSS or highest R^2 value] as M1
    4. In the third step, the model will train with 2 features at a time and select the best result as M2. This goes on till the last step.
    5. In the last step , the model will train with all the features together and save as Mp
    6. Now the model will compare the results from M0 to Mp and save the one with the highest adjusted R^2 value.
    7. Adjusted R^2 value is preferred over R^2 value because the R^2 value always goes up when we add new features to the dataset and hence we will always get Mp as the best model because it has the highest number of features.
4. Forward Stepwise Selection Method

1. Assuming we have p features in the dataset
2. In the first step, the model will train with no feature and simply give the mean value and save as M0
3. In the second step, the model will train with 1 feature at a time and select the best result [lowest RSS or highest R^2 value] and save as M1
4. In the third step, the model will train with 2 features at a time with one feature from M1 and select the best result and save as M2
5. In the fourth step, the model will train with 3 features at a time where two features are the one from M2, select the best result and save as M3. This goes on till the last step
6. In the last step, the model will train with all the features and save the result as Mp
7. Now the model will compare the results from M0 to Mp and save the one with the highest adjusted R^2 value.

5. Backward Stepwise Selection Method
    1. Like forward stepwise selection method.
    2. But here we start from p features and keep on reducing the features at every step till we are left with no features.
    3. This method can’t be used if the number of features is greater than the number of observations.