PROJECT EXPLANATION.

Page:
Date: / /

① Define the business problem.

I need to create a model which will predict the state in which the loan provided to customers will end whether Good or Bad? ie. Good customers - will repay loan properly with EMI while Bad customers won't.

"The machine learning model will predict whether the loan will end up in Good or Bad state?"

② Identify the target variable.

ie. the feature in data you want to predict (bad_loan)

[It is not always easy to identify the target variable since databases are not user friendly & may have technical naming conventions. A Business Analyst can help us in that case]

③ Choosing the appropriate type of ML.

For Continuous variable predictions like Sales, TurnOver, Profit, Demand, Volumes etc. use regression.

For categorical variable predictions like bad_loan = 0/1 use Classification.

④ Remove Useless Variables from Data.

Variables that are useless for algorithm because they cannot hold any patterns with respect to the target variable. eg: Id, Name, email, phone_no.

Business domain knowledge helps in distinguishing useful & useless columns.

Like - 'Age' is continuous numeric variable but it can be important from business perspective.

Also 'date' can be used but deriving it in months because it may hold some pattern.

Page :
Date : / /

Like term < 30 months
60 months.

This variable may be derived from loan-start-emi to loan-end-emi date.

The process of creating new column from existing column is known as Feature Engineering.

(5) Identify 'potential' predictor variables in data.

ie. to identify which factors affect the target variable?

Every business problem is driven by certain factors. For example:

1) What is the number of payments / tenure of loan?
2) What is the purpose of loan?
3) The cibil score of person's account? (deliquency)
4) The amount of loan?
5) Annual income of customer.
6) Customer's financial status identification / security purpose - Whether owns house or no?
7) Interest rate on loan provided?
8) Debt-income ratio of consumer?
9) Whether loan is verified or not?
10) How many people are in customer's family & how many of them are employeed?

How to understand which factors affects target variables?

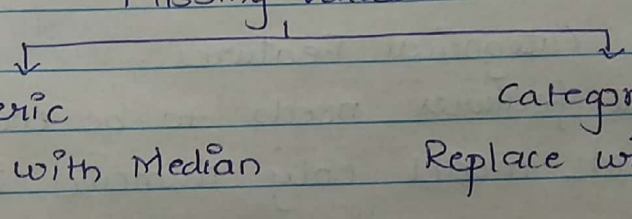
1) Talk to BA or client to get insights.
2) Explore each variable based on how its values are distributed ie. Exploratory Data Analysis.

3) To find distribution of each variable, use histogram or distplot for continuous variables & bar-chart, countplot for categorical variables.

4) The idea is to see distribution of variables/values in a colum. If it is too far away from ideal bell curve that column may not be useful.

⑥ Treatment of missing data in each one of the predictor variable.

Missing values must be removed or replaced, or else it will bias the results produced by ML algorithms.



NOTE:

- If there are more than 30% of missing values in a column, remore that colum.
- If n(missing) << n(total rows) delete rows with missing values. For eg: if there are 5,00,000 rows & 10 rows have missing values. It is safe to drop those rows.

⑦ Treatment of Outliers → identified by Boxplot.

1) Remove outliers from data, But it may cause data loss.

2) Take log() transformation. Apart from that you can also use square root(), cube root, 1/n, exponential transformations which best help to remove outliers.

Page :
Date : / /

⑨ Predictor variables feature Scaling & Spliting of data in train & test set.

→ Independent / Predictor variables needs to be scaled / normalized so that each feature contributes approximately proportionally.

→ Spliting data in train & test after assigning x & y with independent & dependent variables respectively.

→ For Scaling we used, Random Scaler.

→ Train-test-split in 80-20 or 75-25.

⑧ Encoding Categorical Features.

- Categorical features needs to be encoded since for building ML model only numeric inputs are required.
- We used Label Encoding to avoid the cure of dimensionality.

⑩ Creating Model on Training Data.
using ML algorithm as per target variable.

Supervised ML

Regression	Classification
- Linear Regression	- Logistic Regression
- Decision Trees	- Decision Tree
- Random Forests	- Random Forest
- XGBoost	- SVM, Naive Bayes, KNN
- KNN	

Page :
Date : / /

(11) Measuring Accuracy on Testing Data.

- Accuracy is measured by computing few metrics.

Regression → Median Absolute Percent Error (MAPE)
Mean Absolute Percent Error

Classification → Precision, Recall, F1-score, AUC, Confusion Matrix.

- A comparison between all values of is done to see which algorithm is producing best accuracy for given data.

(12) n-fold Cross Validation (Bootstrapping)

- There is a chance that while selecting records for Training data we got neat & clean records which resulted in High Accuracy.
- In order to be sure that this accuracy is consistant all the time we select a random sample multiple times from full data. Then train & test the model on data.
- ie Apply Cross Validation.

(13) If Target Variable is highly imbalanced.

- We use 'imblearn' library.
- Which includes methods like Undersampling, Oversampling, SMOTE, Tomek etc.
- Build the model on new samples obtained and measure accuracy on its test data.

(14) Finding the importance of each predictor statistically.

Which of the used predictors is really affecting the target variable?

- It can measured using:
  - Heatmap,
  - ExtraTreeClassifier
  - SelectKBest Features.

(15) Train the predictive model on full data.
- When you are satisfied with accuracy of model & final set of predictors are selected. It is time to train the predictive model using complete available data.
- It helps in exposing all types of pattern in data.
- If similar pattern encountered in live environment, model will be able to predict answer accurately.

(16) Deploy the predictive model in production.
1.) Save the model as a serialized file which can be stored anywhere (.pkl file)
2) Once serialized model is placed at server, then any front end application can access it whenever required.