Name: __Wen Dong__ Student ID: _110057395_

Speaker Name:__ Alay Majmudar _ Speaker Title: __Associate data scientist ___

Title of Talk: _modelling process_ Date of Talk:_30th July_

Modelling Process

The speaker demonstrated the complete process of building a machine learning model for a real-world business problem, walked audience through in great details each step, from analysis of the business situation to gain understanding of the value of framing it into an AI paradigm, to exploratory data analysis, data pre-processing, feature selection, model selection, hyperparameters tuning and eventually model training. It gave us in-depth insight and preview into the industry-level modelling process and connected our conceptual knowledge with concrete enterprise business process.

Analysis of business problem

Loan application outcome prediction for a house finance company was taken as the example for the demonstration. Just as former speaker cited from Andrew NG “I don’t want to hear about your AI problem, I want to hear about your business problem”, every AI modelling starts with business problem analysis as it is the fundamental element for AI practitioners to understand not only the value of the modelling but also how to model the problem. In this case, by applying modelling, it would significantly expedite loan application process which not only increase customer satisfaction but also reduce the cost and human effort that incurred in the traditional man-power approach and thus relieve limited company resource to reach to a wider range of potential customers.

Exploratory data analysis

The speaker loaded dataset and inspected to each column to learn the data types, whether numeric or string, categorical or continuous, and data distribution among the possible values, whether it is normal or skewed. The correlation between features and label was explored as well to identify important input matrix.

Data pre-processing

It went through missing value pre-processing, outlier detection, encoding and numerical data normalization and standardization.

1. Missing value

Missing value is inevitable in dataset from real world, since computer is unable to tackle null or empty values, it must be eliminated either by dropping records with missing values or by filling it by imputing. Several strategies can be adopted in imputing, a) treat missing values as another value; b) treat it as a special category in categorical feature; c) fill median or mean value in a numerical field; d) impute with ML algorithms like KNN or linear model.

Outlier detection aims to detect outliers in certain data columns and contain them, as outlier is abnormal data that might lead to fool the model into an incorrect direction. It can be detected with box plot of Inter Quartile Range. Several measures can be implemented for that, a) completely delete outliers; b) transform them into inliers; c) imputation or treat them differently, among those options completely deleting is sometimes the best.

2. Encoding

It consists of categorical encoding and numerical encoding. Categorical encoding is to encode categorical text strings into discrete numbers in reasonable ways like a) one-hot encoding; b) binary encoding, which encodes into binary format of number; c) base-n encoding, which is special case of binary encoding, binary is 2 based while base-n is integer n based. Numerical encoding is to scale numerical data into a normal or standardized distribution, which can remove unnecessary advantage or influence due to distinct distribution of features. Min-max and standardization scaling are the mostly used.

3. Sampling

Sampling is to ensure all labels are attended and represented especially the minorities during a training cycle. Image if some minor labels are missing during the training, how the model can predict such label during testing or validation. Stratified sampling, down sampling, up sampling, and hybrid sampling are the usual procedures to address this problem.

4. Pipeline

Pipeline is to chain the pre-processing phases so that the data is flowing through the sae pre-defined processing phases from source batch by batch into model training process, it is especially useful when the source data is too large to load at once.

Feature selection

Feature selection is to identify relevant columns that related to output labels, usually dimensions reduction by removing correlations or highly sparse features is the way to go.

Model selection

Feed the same training dataset into different machine learning algorithms to get identify the best-resulted model. As different algorithms by their nature might perform distinctly over different dataset, simple way is to feed the dataset into a set of selected models and observe the performance. To avoid overfitting, cross validation could be applied. It would be great it if the speaker could clarify in more dimensions on how cross validation help tunning hyperparameters.

Hyperparameters tunning

Hyperparameters are the parameters irrelevant to dataset itself but related to model structure or training process, like learning rate, epoch size, et cetera. The selection of hyperparameters has significant impact to the model performance. It came the running, there are optimization tools like sk-optimize and Keras tuner can help running, which basically try different combinations of hyperparameters and pursue the optimal.

Conclude

Overall, the speaking was very informative and helpful to transform our academic view to the industrial perspective, articulated almost every single detail during the modelling process.