ShinyPET: A Predictive, Exploratory and Text RShiny Application using Airbnb data

Ang Su Yiin
Singapore Management University
suyiin.ang.2020@mitb.smu.edu.sg

Joey Chua
Singapore Management University
joey.chua.2020@mitb.smu.edu.sg

Kevin Gunawan Albindo
Singapore Management University
kgalbindo.2019@mitb.smu.edu.sg

ABSTRACT

The increasing availability of data has resulted in the increased demand for data driven decisions. Although there is an extensive range of commercial statistical tools, they are often subscription-based and demand good technical knowledge to mine and draw insights from. Therefore, it may not appeal to the average user.

As such, our project aims to develop a user-friendly application that will enable users to make data-driven decisions without the need to understand programming languages or have extensive statistical knowledge. We will use Airbnb data as our baseline for this project - data generated is rich in information, which consists of structured, unstructured (textual), and location data.

With this application, users will be able to perform text analysis on review and listing data to generate more quantitative insights. The exploratory module allows users to identify interesting patterns based on selected variables. Findings from the exploratory module will be further augmented in the confirmatory module where selection of statistical methods will be guided based on user's chosen variables. Finally, the predictive module enables users to prepare and build a variety of prediction models without needing to have in-depth understanding of the predictive models and its algorithms.

1. MOTIVATION OF THE APPLICATION

xx

2. LITERATURE REVIEW

Review and critic on past works

3. DESIGN FRAMEWORK

A detail description of the design principles used and data visualisation elements built

3.1 Exploratory module

3.2 Text module

3.3 Predictive module

Predictive analytics attempts to make prediction about a specific outcome based on historical data. It is done by understanding the relationship between variables/predictors to develop a prediction model. In general, predictive analytics comprises several steps: data pre-processing, feature selection, model training, model tuning, validation, and model selection.

Tidymodels has become the popular choice in R community which provides a framework for predictive modeling and machine learning. It is aligned with the tidyverse principles which leads to a tidier and consistent grammar in the predictive analytics process. Our predictive module design framework follow tidymodels framework for data sampling, model training, tuning, and validation. On top of that, data pre-processing and feature selection are supported by other R packages such as tidyverse, ranger and Boruta.

To support predictive analytics, visualisation and interactivity are embedded throughout each step:
a. Data sampling - Selection of training-test split proportion provides flexibility in deciding how to spend data budget on the model development process. The distribution plot between training and test set highlights any potential bias in the training data set.
b. Feature selection - Correlation matrix with customised correlation type and p-value criteria, as well as variable importance allow assessment of correlation among variables.
c. Data transformation - Distribution plot between pre and post processing step increases user awareness on what transformation steps are performed and on which variables.
d. Model training and validation - Rsquare plot to visualise training/validation result along with table of metric performance and coefficient estimate or decision tree information as interactive plot to improve result evaluation.
e. Prediction error assessment - Training set distribution plot overlapped with predicted values to allow further investigation on prediction error.
f. Hyper-parameter tuning - Plot of model performance using different hyper-parameters setting helps user to understand the change in performance.
g. Model evaluation - Plot of performance from different models to support clearer model evaluation process.

4. DEMONSTRATION

- use case

5. DISCUSSION

What has the audience learned from your work? What new insights or practices has your system enabled? A full blown user study is not expected, but informal observations of use that help evaluate your system are encouraged.

6. FUTURE WORK

The current predictive module is limited to 5 types of predictive model: linear regression, generalised linear model, decision tree, random forest, and boosted tree. In future, more predictive models can be added to the list, such as neural network (neuralnet / keras package are supported by Tidymodels) to provide user with wider model selection. In terms of hyper-parameter tuning, more parameters can be made available for user input to provide more flexibility in developing predictive model. In-depth statistical analysis in model training such as residual analysis are currently not available and this would be a good additional tool to improve our predictive module.