TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

Nitesh Kumar Prajapati

Agenda

1. Data Exploration
2. Model Development
3. Interpretation

Data Exploration

Understand the characteristics of given fields in the underlying data such as variable distributions, whether the dataset is skewed towards a certain demographic, and the data validity of the fields. For example, a training dataset may be highly skewed towards the younger age bracket. If so, how will this impact your results when using it to predict the remaining customer base? Identify limitations surrounding the data and gather external data which may be useful for modeling purposes. This may include bringing in ABS data at different geographic levels and creating additional features for the model. For example, the geographic remoteness of different postcodes may be used as an indicator of proximity to consider whether a customer is in need of a bike to ride to work. Exploration of interactions between different variables through correlation analysis and look out for multicollinearity by creating interaction variables. An example of this correlation may occur between independent variables age and tenure – i.e. people of the older brackets will have a longer tenure. Furthermore, the transformation of required data so that it is in an appropriate format for analysis. This may include steps such as ensuring that the data types are appropriate and rolling data up to an aggregated level. Or, joining in already aggregated ABS data at a geographic level to create additional variables. Document assumptions, limitations, and exclusions for the data; as well as how you would further improve in the next stage if there was additional time to address assumptions and remove limitations.

Model Development

Determine a hypothesis related to the business question that can be answered with the data. Perform statistical testing to determine if the hypothesis is valid or not. Create calculated fields based on existing data, for example, convert the D.O.B into an age bracket. Other fields that may be engineered include 'High Margin Product' which may be an indicator of whether the product purchased by the customer is in a high margin category in the past three months based on the fields 'list_price' and 'standard cost'. Other examples include calculating the distance from the office to the home address as a factor in determining whether customers may purchase a bicycle for transportation purposes. Additionally, this may include thoughts around determining what the predicted variable actually is. For example, are results predicted in ordinal buckets, nominal, binary, or continuous? Test the performance of the model using factors relevant to the given model chosen (i.e. residual deviance, AIC, ROC curves, R Squared). Appropriately document model performance, assumptions, and limitations.

Interpretation

Visualization and presentation of findings. This may involve interpreting the significant variables and co-efficient from a business perspective. These slides should tell a compelling storing about the business issue and support your case with quantitative and qualitative observations. Please refer to the module below for further details

THANK YOU