TATA CONSULTANCY SERVICES

Building on belief

White Paper

Feature Engineering

Key to Maximizing Business Outcomes with Predictive AI

Introduction

Retailers worldwide are leveraging AI/ML technologies to derive predictive insights and respond to market dynamics in real time. Every intelligent system used by retailers has at its core machine learning, deep learning, image/video processing algorithms or other statistical models. These models are more or less similar in structure; the success of any AI/ML model depends on how well the raw data sources are processed to extract key features that are the flag bearers of retailing. Feature engineering in data science is the process of transforming (aggregating or decomposing) raw data into key features of the data that can be used as input variables to get the best results from algorithms.

This paper explains how to isolate key information from data noise, connect the dots, and highlight patterns to maximize the outcomes from AI implementations with feature engineering, and get a clear edge over competition.

Myth
Advanced algorithms are the panacea for AI inaccuracies.

Reality
The quality of input data and the features derived therefrom determine AI outcomes.

Myth
AI is all about science.

Reality
AI success requires an interplay of the art of applying domain and technical expertise with data science to address business problems.

Feature Engineering: The Art of Finding the Diamonds in the Data

Isolating key information and highlighting patterns encompass a blend of domain, analytical, and data processing expertise. Let us understand the six key steps to effective feature engineering with an example of macrospace optimization (see Figure 1).

1. Identify the Data Sources

Space allocation in a retail store is influenced by many factors that may have either a positive or negative impact on sales. For example, sales, margin, units, demographics, promotion, competitor, competitor strength, local weather, inventory, and labor cost.

The first step in feature engineering is to identify all the factors and then gather and process it. Each parameter may require different levels of processing.

For example, store-specific demographics can be gathered based on the store location and demographic data available at the zip code level (typically obtained from third-party vendors). A single store could cater to many zip codes under a trade area, in which case the demographic variables of the relevant zip codes need to be processed to derive the store-specific demographic variable. In most cases, it is assumed that all categories will have the same trade area and the demographic variable is derived accordingly.

Feature engineering for store specific demographic information generation Feature engineering for store specific competitor intensity derivation

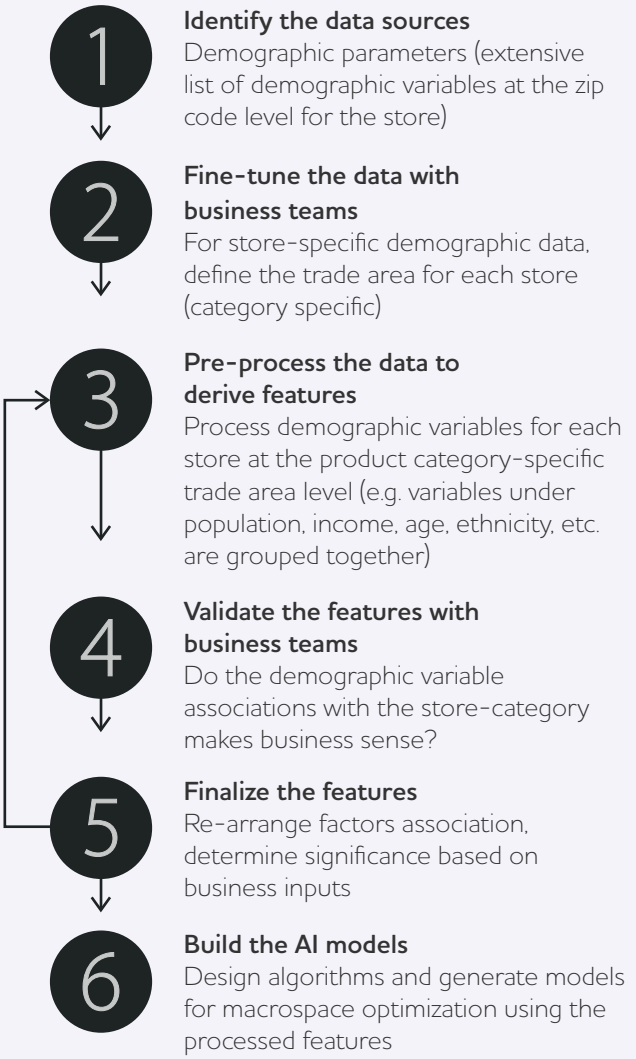

Figure 1: Feature engineering for store-specific demographic information generation

- Let us consider another example that measures the intensity of competition. Based on the inputs gathered from business, 'N' competitor stores were identified for a given retail store and their average distance is derived. These derived features would define the competitiveness of other retailers in the trade area.

2. Fine-Tune the Data with Business Stakeholders

Work with business stakeholders to group the identified data parameters and understand their impact on business outcomes. This is a critical step.

- Continuing with the example of store-specific demographic generation, inputs from business stakeholders are leveraged to define the trade area for each store. The trade area is highly dependent on the product category. For example, a customer might not be willing to travel more than two miles to buy dairy products such as milk but might be willing to travel even up to 10 miles to buy an electronic item such as TV. Zip codes within two miles and 10 miles comprise the trade areas for the diary and electronics categories, respectively. And the demographic variables for the category are processed based on the zip codes and category characteristics.

- The trade area is identified for each store (by category) as explained above. Also, based on the inputs gathered from the business, major competitors are identified for the categories and ranked on a scale of 1 to 10, to measure the store-category level competition strength.

3. Pre-process the Data to Derive Features

Next, apply statistical models to validate the impact of data parameters on business. The prime objective is to derive features, establish relationships, and reduce the number of data parameters.

- Going back to the example of store-specific demographic generation, demographic variables (such as total population and average per capita income) are available for each store by category and trade area. These are processed using the variable reduction technique to reduce the number of features, without much information loss. It is important to strike a balance between too many and too few features. Too much information might create ambiguity in explaining the same behavior, as some variables might be strongly co-related. For example, variables pertaining to a trade area such as male population, female population, total population, number of households, population density, and similar variables are giving population information in different aspects; all these variables need to be condensed into a single feature without losing much information using suitable statistical techniques such as principal component analysis (PCA). Similarly, variables related with income are condensed into another feature and so on.

- Based on the competitor distance and competition strength, competition intensity is derived at the individual category level for a retail store. Competition intensity depicts the combined impact of competitor distance and competition strength.

4. Validate the Features with Business Stakeholders

Next, validate the established relationships with business stakeholders. Highlight any anomalies observed to modify the features identified at the initial step.

- For example, we have distributed the demography in a trade area between the two stores (A and B) based on the trade area distance. Out of the 1000 customers, 500 belongs to store A and 500 belongs to store B for category A; whereas 250 belongs to store A and 750 belongs to store B for category B.. We need to validate this calculation with the business.

- When analyzing the relationship between performance and competition, it was identified that the relation was not always true. There were multiple instances where a store performed well even when the competition strength was high. Based on the business inputs, it was identified that along with the competition strength, the competition's impact is also based on the product category and the distance between competitor stores. These interventions from the business, supported by the findings from data analysis would prune the process further. For example, a general merchandise retailer would consider an electronics specialty retailer as a strong competitor for their electronics category and a pet specialty retailer as a strong competitor for their pets category, but not vice-versa.

5. Fine-Tune and Finalize the Features

Based on the validation from the business and additional findings, data pre-processing and business validation steps need to be reiterated to finalize the feature set. In addition, define strong differentiating features by transforming numerical variables into categorical variables. Also, derive and assign weights to features based on the business considerations.

- For category A, the population was equally distributed between stores A and B; however, for category B, it was distributed in the ratio of 25:75. In this step, we distribute the complete population between stores for a category.

- Thus, it was concluded that competition is inversely related to competitor distance; weights were derived by considering the rank of competitor and the category-wise competitiveness. Additionally, competitor intensity varied based on the season. For example, the lack of transport facilities and adverse weather conditions during winter impacted the competitive rankings for each category.

6. Build the AI Models

Finally, integrate all the identified features to design algorithms and generate models for solving the business problem (macrospace optimization, in this case).

The steps outlined above are the recommended practices. Determining the appropriate feature engineering steps/process for a given context is an iterative process that culminates with attaining the maximum accuracy feasible with the available data.

Automation of feature engineering should be undertaken only when the feature engineering process and the predictions of the AI models mature, i.e. the results should be consistent over a period of time with maximum accuracy for a specific business need such as macrospace optimization. Any automation is suitable only for the specific business need of the retailer and it may not be applicable to other business areas of the same retailer or even same business area of other retailers.

Conclusion

Algorithm selection and model training without doubt determines the success of AI in retail; but the most critical piece is not how we select or tune algorithms but what we input to AI/ML, i.e., feature engineering. Better features provide the flexibility to use simpler models that are faster to run and easier to understand and maintain.

About the Authors

Jeisobers T, Principal Scientist, TCS Optumera
Jeisobers has 11 years of experience in TCS and 19 years of overall experience in advanced analytics. He conducts research in merchandising areas including macrospace, assortment, demand transfer and pricing, and contributes to intellectual property through patent generation.

Pranoy Hari, Retail Functional Consultant, TCS
Pranoy Hari has over nine years of experience in space planning, assortment, planogram, pricing, and promotions. He has spearheaded multiple business consulting assignments for leading global retailers in the merchandising domain. He specializes in merchandising strategy, retail digital innovation, and business transformation practices.

TATA CONSULTANCY SERVICES

Building on belief

About TCS Retail

TCS Retail partners with over 100 global retailers, driving their growth and digital transformation journeys. We are solving their toughest challenges by harnessing our deep consulting and technology expertise, amplified by strategic investments in products and platforms and research partnerships with top universities; a co-innovation ecosystem of over 3,000 startups; and Nucleus, our in-house innovation lab.

Retailers worldwide are adopting the TCS Algo Retail™ framework, a playbook for integrating data and algorithms across the retail value chain, thereby unlocking exponential value. Our solutions and offerings leverage the combinatorial power of new-age technologies to make businesses intelligent, responsive, and agile.
TCS' portfolio of innovative products and platforms include AI-powered retail optimization suite TCS Optumera™, unified commerce platform TCS OmniStore™, and AI-powered enterprise personalization solution TCS Optunique™. With a global team of 40,000 associates, we are powering the growth and transformation journeys of leading retailers worldwide.

Contact

For more information on TCS' Retail Solutions and Services,
please visit http://on.tcs.com/TCS-Retail

Email: algo.retail@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is a purpose-led transformation partner to many of the world's largest businesses. For more than 50 years, it has been collaborating with clients and communities to build a greater future through innovation and collective knowledge. TCS offers an integrated portfolio of cognitive powered business, technology, and engineering services and solutions. The company's 469,000 consultants in 46 countries help empower individuals, enterprises, and societies to build on belief.

Visit www.tcs.com and follow TCS news @TCS_News

Corporate Marketing | Design Services | M | 04 | 21