Predicting Sales Volume of Pet Products using Product Title Keywords

Evidence from Chinese E-Commerce Platforms

Nicole Neo, Naijia Wu, Jackie Zhu, Lin Zhu
(Group 4)

Overview

- Background
- Data Description
- Methodology
- Results
- Conclusion
- Reflection

Background

- Research question: Which are the top 10 keywords in the names of pet products with the highest sales on e-commerce platforms in China?
- Hypothesis: Words like “organic”, “natural” and “imported” are among the top 10 keywords in the names of pet products with the highest sales.

Data Description

- Product sales data from Taobao and Jingdong (e-commerce platforms in China)
  - Timeframe: **Oct 17 2017 - Oct 22 2017**
  - 703 500 observations
  - json format
  - Main fields of interest: pro_sales_num, pro_class, pro_name

Methodology

Pre-processing

Data exploration
- Choose fields to analyze
- Filter subcategories

Text cleaning
- Remove English/Mandarin punctuation and digits

Tokenization
- Using jieba (suitable for Mandarin tokenization)

Vectorization

CountVectorizer

TF-IDF

Modelling, Validation, Evaluation

Models
- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest

Data
- X: Vectorized product title data
- Y: Sales volume

GridSearchCV
- Default 5-fold CV with R-squared scoring
- Hyperparameter tuning

Evaluation
- Select model with best test score + least overfitting
- Examine top coefficients

Results

Model Selection

- Based on the test set score that represents the model ability to generalize to new unseen data and the overfitting problem, we finally choose Lasso Regression with TF-IDF data

Model with CounterVectorizer Data

Model	Training set score	Test set score
Linear Regression	0.998	negative
Lasso Regression	0.799	0.510
Ridge	0.889	0.554
Random Forest Regression	0.696	0.536

Model with TF-IDFData

Model	Training set score	Test set score
Linear Regression	0.996	negative
Lasso Regression	0.913	0.735
Support Vector Regression	0.922	0.707
Random Forest Regression	0.731	0.583

Results

Keywords interpretation:

- By ranking the coefficients from the Lasso Regression model based on absolute value, we get the most important keywords to the pet products' sales

- Among the top 40 most important words to facilitate the sales volume, we get some representative words including **“干无盐”(”dry and no salt”), “多种”(“multiple”), “包邮”(“free deliver”), “标准”(“Standard”), “冰淇淋”(“ice cream”), “鸟粮”(“birdseed”), “秋田”(“Akita”), “躲藏”(“hide”) , “警示灯”(“warning light”), some brand names, etc**

Conclusion

Application

- NLP + Marketing & Product Naming
  - ➔ Help address online marketing directions
  - ➔ Give top product name tokens
  - ➔ Provide information about features that make a product succeed
- NLP + Business Analysis
  - ➔ Predict sales volume for given product titles
  - ➔ Allow use to plan for demand
  - ➔ Help increase their profits

Reflection

- Insufficient data and lack of adequate historical sales trends assessment
  - ➔ Our dataset only contains daily sales and product data from Taobao from 10/17/2017 to 10/22/2017
  - ➔ The market is a dynamic place
- Model accuracy and runtime trade off
  - ➔ Model accuracy and runtime are two important factors to be considered during the model training process
  - ➔ Our eventual Random Forest Regression model did not perform as well compared to Lasso Regression.

Thanks!