Indium Software uses Machine Learning-Powered Product categorization to increase conversion rates

Situation

The client is an AI-powered e-commerce aggregator website which delights customers by providing them with smart buying options. The product categorization is indispensable for e-commerce websites. It makes free-text searches faster and provides better user experience by highlighting top categories upfront.

While this categorization works well on retailer websites, it becomes an issue for e-commerce aggregators. The product categories are defined differently by different retailers for the same product. This creates a problem in assigning the same product from different retailers to the same categories due to which the quality of search results and user experience suffer.

There was a need for Advanced Machine Learning and Artificial Intelligence techniques to be deployed to solve some of the most complex problems the industry faces. In particular, this would be a case on how we solved the nagging problem of product categorization.

Challenges

Availability of training data – ML algorithms need training data to learn the right answers before they start predicting. The data from historical business processes was unavailable. We could have created a training data manually or through some heuristic method. The manual approach is time-consuming and error prone. We used a rule-based heuristics application.

Large scale data - Product listings for e-commerce aggregators are hundreds of millions. Such huge data requires huge storage and distributed computing power. The text classification needs text pre-processing methods which are themselves very computation intensive. We used a 32GB RAM machine on AWS for prototyping the model and running experiments.

Solution

Here is a summary of the steps involved and the methods used in each step of this process.

- **Data Sampling**
  - To get equal representation from each category
  - Methods : Stratified Random Sampling
- **Pre Processing**
  - Convert Text to a numeric representation
  - Methods : TF-IDF ,N grams, Stop words, stemming lemmatization
- **Model Training**
  - Fit classification models to training data
  - Naive Bayes, Support Vector Machines
- **Parameter Tuning**
  - Find model parameters that give the highest accuracy
  - Grid Search (Scikit-learn) , Cross Validation
- **Model Nesting**
  - Train separate models for different levels of hierarchical groups
  - Group By methods, For-loops, Sub-setting
- **Production**
  - Setting up infrastructure for the trained model to be utilized
  - Pickling, Django, JavaScript

Tools Used

Python, Scikit Learn, Jupyter Notebook (for prototyping) PySpark, Pickling using Django/Flask (for scaling and productionising)

Machines used

Windows 8GB RAM (for prototyping)
AWS 128GB EC2 machines (for scaling)

Engagement Model

Offshore

Business Impact

- Model trained achieved an accuracy of **75 %** in accurately predicting categories for new products
- Improved Product Categorization lead to superior indexing of products which directly contributed towards providing better search results
- **3 % Increase in Conversion rates** across all Product Categories which led to a **20% increase in GMV**

About Indium Software

Indium Software is a Big Data Technology Consulting and Professional Services company. Indium Software focuses exclusively on Global Enterprises and help them leverage Big Data through Strategic Consulting, Technology Selection, Design & Architecture and Implementation of the solution. With offices in the U.S and India, Indium Software offers Flexible Engagement Models and Comprehensive Solution Approach to help customers leverage Big Data

For more information, contact: **info@indiumsoft.com**