One-page summary of industry lecture 2 (LTK)

Shaun Pritchard

Florida Atlantic University

CAP 5625

October -20-2021

M.DeGiorgio

liketoknow.it (LTK) is an enterprise affiliate shopping discovery app that allows consumers to shop for clothes, makeup, home décor, and other products shown by their favorite social media influencers. Data science and technology platforms and practices that drive LTK's proprietary applications are discussed in this lecture. LTK generates $2.9B in annual sales by bringing together more than 5,000 brands, 150,000 creators, and 8M shoppers through original content created by its community of influencers and creators.

The lecture discusses the proprietary technologies and the backend for LTK's enterprise development. I enjoyed learning about the technologies used in LTK from the perspective of data science and big data. LTK discussed how to build, train, and deploy machine learning models on the cloud using Sagemaker from Amazon. using GitHub for Version Control, docker to containerize code base modules and applications.

They describe the backend data environment and architecture of the LTK and the big data foundation it has. The LTK machine learning life cycle involves processing, training, testing, and deploying data to the LTK machine learning models repositories, which work on contextual and image-based datacast services for recommendation systems Trend forecasting, and product classification.

As a result of exploring the groundwork for retail production classification (RPC), which is the core of their project, I can help accurately classify products on their application that influencers post every day, using machine learning to leverage and understand the integral aspect of their product catalog and trends.

LTK RPC implements two different machine-learning models together, namely an image model and a text-based model. The image model made use of a ResNet50 base model with 1024 nodes, whereas the text model makes use of two dense neural network layers with 256 nodes, where all neurons in each layer are connected to all neurons in its preceding layer. Based on a count of each word in the string, they turn the text data into a vector, then take the image information within the text data to develop product categories.

In addition, they use over 50,000 manually categorized products to predict the category from their data lakes using both texts from the title and description of the image. In terms of machine learning, they use a categorical cross-entropy classifier with one-hot encoding for labels.

In order to solve the problem of class imbalance, they use oversampling methods which consist of redundancy for images, image augmentation, SMOTE for text data, and under-sampling to randomly remove instances from the majority classes.

RPC is tuned using hyperparameters such as drop out, where nodes are randomly removed to prevent overfitting, batch size, learning rate alpha, max count value determines the number of instances in each class after solving the class, and balance, and weight decay. The image augmentation parameter is a rotation range, a brightness range, a sure range, a zoom range, a channel shift range, and a number of layers.

Then we examine the architecture for the combined efforts across data science, data operations, and data engineering. Creating new product domains in the data lake, creating flexible machine learning models that accurately categorize products, and creating the infrastructure for deploying machine learning models at scale to support millions of automated records.

I also found it uniquely interesting that their big data architecture combined with artificial intelligence has yielded 11 million global products categorized, 45 million influencer products categorized, a 92% Gross metricized value when an influencer gets a sell what value they get for it and 94% of their Global and Link products categorized all through the implementations of AI in Big Data.

In addition, they mentioned that creating training and building models was an easy aspect of their implementation. To achieve the results they have produced with their hard work, it was essentially getting all these different technologies to talk to each other and communicate correctly.

This lecture gave me a glimpse into how a successful tech company applies what I've only been practicing in theory. Moreover, I gained more insight into using big data for text and image classification. Even though I wish I had the opportunity to ask them some questions, I found the lecture very powerful.