BI
Summer
Internship

Virtual Intern Presentation

Session #7

Wenxin Zhang

Boehringer
Ingelheim

PRESENTER BIO

- *Name: Wenxin Zhang*
- *Major: Data Science*
- *Function & Role: Data Scientist Intern*
- *College/University: New York University*
- *Graduation Date: 2022 May*
- *Fun Fact About Self:*
  - *I love all kinds of sports (dance, swim, badminton, ping-pong, basketball…)*
  - *I come from Sichuan, a beautiful province in China, famous for hot-pot and Pandas.*

Project1-Build Dashboard for Sales Forecast

Project Description

In terms of the pet products including Flea & Tick, and Heartworm; the pet products brand including Nexgard, Frontline, and Heartworm, build dashboard to visualize the predicted future sales on both brand level and SKU level

Goal of Project

Obj 1: Preprocess data & Understand and apply feature selection model (lasso regression) and SARIMAX time-series prediction model

Obj 2: Empower the short-term and long-term models' results to build dashboard for visualizing predicted sales

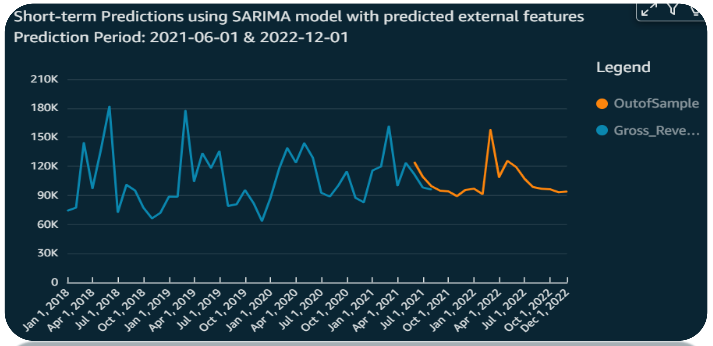
Obj 3: Advice finance group on revenue prediction and budget setting; advice supply chain group on manufacture planning

Current Progress

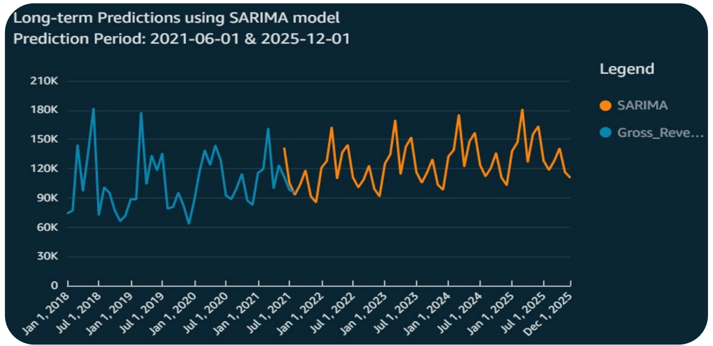
Preprocess the sales data, collaborate to select features and make predictions using time-series model

Visualize short-term and long-term sales prediction at SKU level and brand level using line plots, pie charts, box plots…

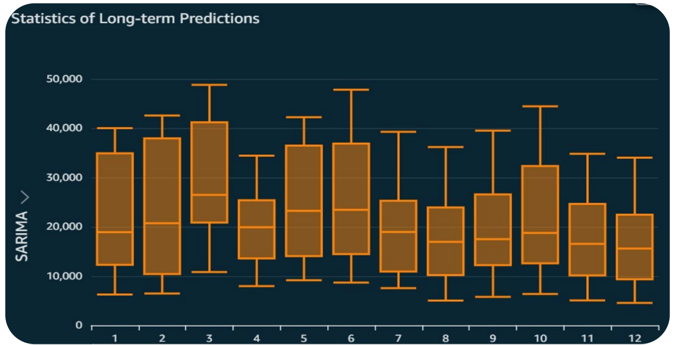
- Add 2 filters
  - 1) Brand (e.g. Heartgard)
  - 2) Time Duration (e.g. 2021-06-01~2022-12-01)
- Prediction at brand level



Short-term pred 2021-6-1~2022-12-1

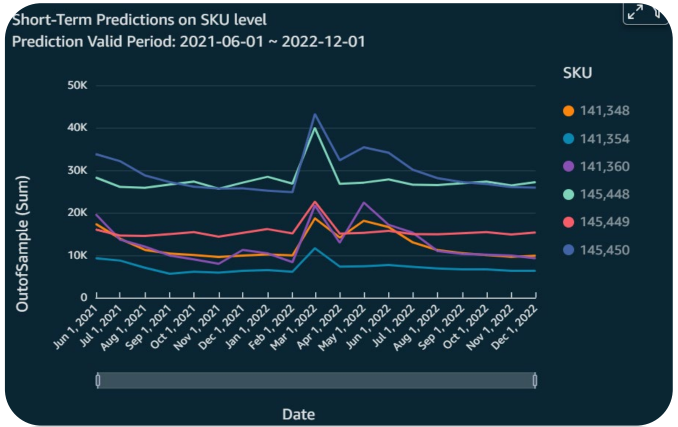


Long-term pred 2021-6-1~2025-12-1

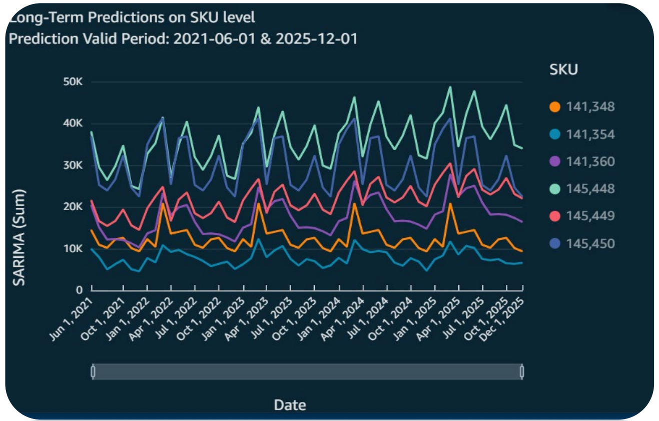


Box Plots of long-term pred

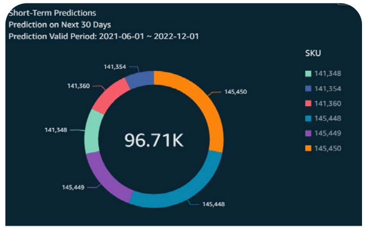
- Prediction at SKU level

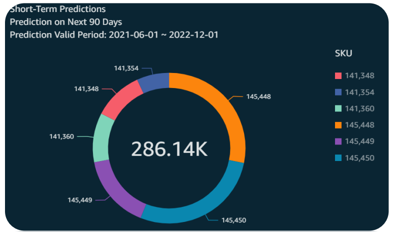


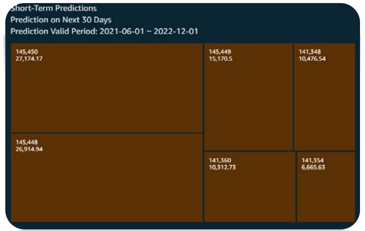
Short-term pred
2021-7-1~2022-12-1

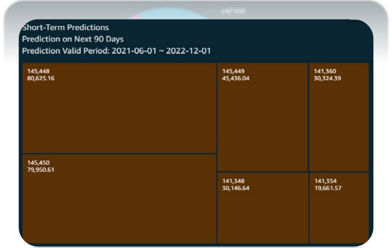


Long-term pred
2021-6-1~2025-12-1









Project2-Cluster Deviations

Project Description:

- Deviations include all kinds or reported problems of BI, such as the broken manufacturing pipelines, the temperature shift, and the Covid-19 detection. Clustering these deviations can help detect the trending of these problems, and provide informative patterns when problems reoccur.

Goal of Project

- Obj 1: Preprocess the multi-lingual text data
- Obj 2: Cluster the deviations using K-Means, and tune parameter(number of clusters)
- Obj 3: Cluster the deviations using HDBSCAN, and tune parameter(metric, min samples per cluster, algorithm)
- Obj4: Check what kind of descriptions(e.g. short description, detailed description, summary description) work best for clustering

Current Progress/Result

- Preprocess, tokenize the multi-lingual text data using NLTK, SpaCy, and BertTokenizer
- Vectorize the multi-lingual text data using N-grams, Tf-Idf
- Find out when the number of cluster equal to 95, K-Means get best performance, but the situation changes with more data coming in…
- Cluster the data into less than 5 groups, where there are clear patterns.
- Rudece the granularity, extract clear “labels” for each cluster using topic modeling method LDA (e.g. temperature, covid-19)

THIS SUMMER, I LEARNED...

Soft Skills

- Realize importance of being a responsive employee
- Collaborate & Learn from colleagues and other interns
- Move fast & Get used to “learning by doing”
- Be transparent & Grow from mistakes

Hard Skills

- AWS Auto-ML techniques
- Time-series prediction
- AWS Quicksight application
- Multi-lingual Text Tokenization
- Text Clustering

Sincere gratitude to my manager Mariann, my mentor Haritha, Tong and Kriti, and the CDS team!!!