Capstone Project

Rodolfo Jerônimo Teles

11th May 2021

1 Domain background

The domain background chose to this project was the regression problem. Specifically, this project intends to predict future values, according to historical patterns which will be learned by the machine learning algorithm during training process. This study field is known as time series forecasting.

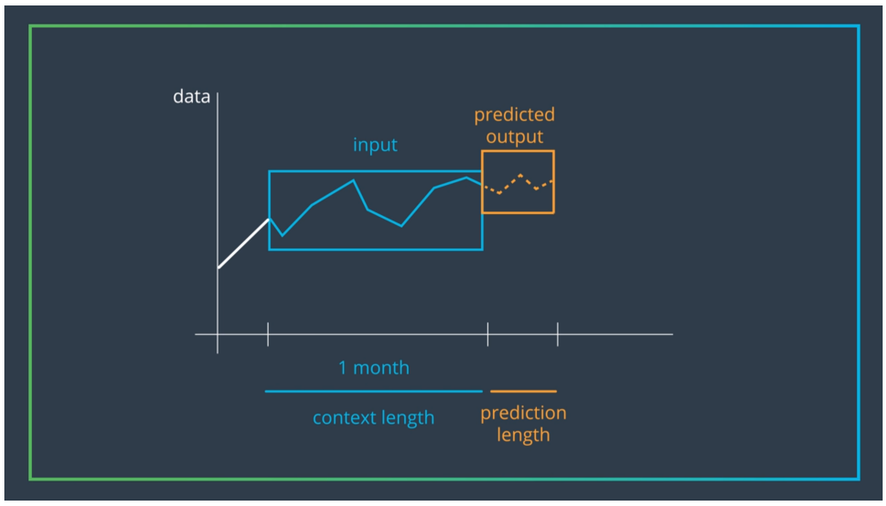

Figure 1: Time series forecasting problem illustration. [5]

2 Problem statement

The Starbucks was founded by Jerry Baldwin, Gordon Bowker, andZev Siegl in 1971 in Seattle (USA) [1] and today is the largest coffeehouse chain in the world. The first app launched by Starbucks was in September 23, 2009 in the App Store. The first version of MyStarbucks application offered a store locator, nutrition information and an interactive drink builder.[2]

Figure 2: Startbucks Logo

The machine learning model used to predict how much a user will spend next day will be deployed using Amazon SageMaker. To be consumed, the model will use a web-app that will allow a user pass his ID, using it as a trigger to Amazon API Gateway which will trigger a lambda function to preprocess the input data and send preprocessed input data to the model. The main tasks involved to create the recommendation system are the following:

1. Exploring and cleaning the data;
2. Choosing a model;
3. Training an estimator that can predict how much someone will spent according to the offer received;
4. Building an API using Amazon API Gateway and make the estimator run on a web-app;
5. Make the web-app predicts a how much a user will spend in each offer to the next day.

3 Data sets and inputs

To train the estimator it will be used three data sets provided by Udacity which contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. The available data are:

1. Portfolio: contains Starbucks offers data. Data available:
    - id (string) - offer id;
    - channels (list of strings);
    - difficulty (int) - minimum required spend to complete an offer;
    - duration (int) - time for offer to be open, in hours;
    - offer_type (string) - type of offer i.e. BOGO, discount, informational;
    - reward (int) - reward given for completing an offer;
2. Profile: contains anonymized users demographics data. Data available:
    - age (int) - age of the customer;
    - became_member_on (int) - date when customer created an app account;
    - gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F);
    - id (str) - customer id;
    - income (float) - customer's income.
3. Transcript: contains data about users actions related on offers like: users interactions, offer received and offer viewed.
    - event (str) - record description (i.e. users interactions, offer received, offer viewed, etc.);
    - person (str) - customer id;
    - time (int) - time in hours since start of test. The data begins at time t=0;
    - value - (dict of strings) - either an offer id or transaction amount depending on the record.

4 Solution statement

The solution consists in using Amazon SageMaker's supervised learning model, DeepAR, to predict how much someone will spend based in the next days. Knowing that information, it is possible to discover which demographic groups respond best to which offer type and create a recommendation system that suggests customized offers .

According to DeepAR forecasting algorithm documentation, it is a method for finding time-based patterns and forecasting scalar (one-dimensional) time series using recurrent neural networks(RNN).

5 Benchmark model

A great benchmark model to compare to the expected solution is the model used during the Time Series Forecasting class in the Machine Learning Engineer nanodegree from Udacity.

The Udacity's case study consisted of use household electric power consumption historical data over the globe to predict energy consumption the first 30 daysof 2010. The dataset was originally taken from Kaggle and represents power consumption collected over several years from 2006 to 2010. The Amazon Sagemaker DeepAR model was used to find patterns in historical data and predicting energy consumption. After the model training process the following metrics were achieved in the evaluation step in the test set:

- RMSE: 0.3479;
- mean_wQuantileLoss: 0.1427.

To illustrate the confidence interval plotted according to model predictions in the test set see the figure bellow:

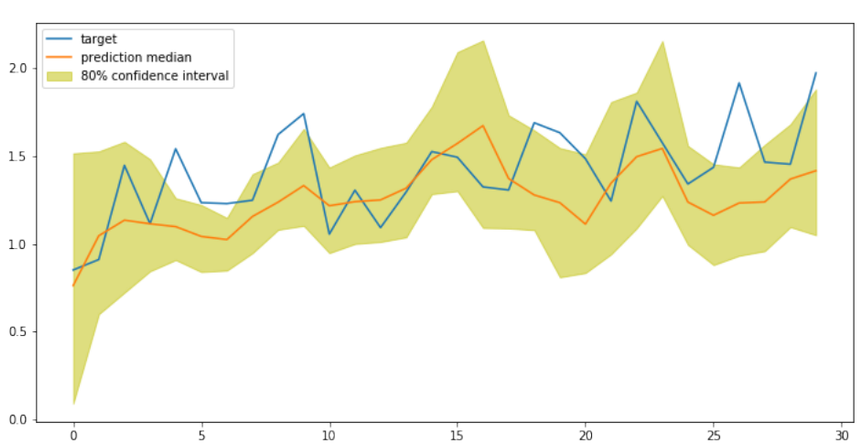

Figure 3: Predictions to energy consumption problem in the test set using Amazon DeepAR model [4].

6 Evaluation metrics

Evaluate model performance is a essential step to get the best predictions as possible to achieve. Thus, to evaluate how well the model predicts the users purchases, it will be used the mean squared logarithmic error (MSLE) which is defined by:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} (log(y_i + 1) - log(\hat{y}_i + 1))^2 \tag{1}$$

Another relevant metric to evaluate the model performance which will be used in this project is the Root Mean Squared Error (RMSE) that is defined by:

$$L(y, \hat{y}) = \sum_{i=1}^{N} \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}} \tag{2}$$

7 Project design

The project design chosen will be using offers, transaction and demographic historical data provided by Udacity to predict the next day total sum value of the user interactions categorized as "transaction". Using the "amount" from the column "value" in the data set "transcript.json" as target, a DeepAR model will be trained passing offer, demographic and users interactions with the app as input of the model.

It will be used the same project design used during sentiment analysis using Amazon SageMaker project in the Machine Learning Engineer nanodegree from Udacity. The design can be viewed at the image bellow:

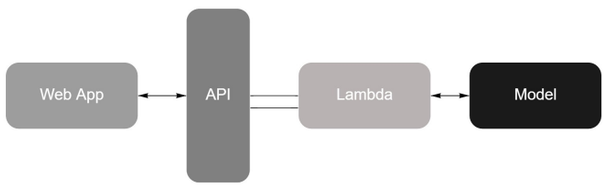

Figure 4: Project design in high level. [3]

Bellow, it was mapped all the steps required to get a deployed model and get predictions using the web app. The process was spplited in five categories:

1. Get and explore the data;
2. Preprocess and upload to S3;
3. Training and deploy a predictor;
4. Pre-process the predictions;
5. Setting support services and running the web app.

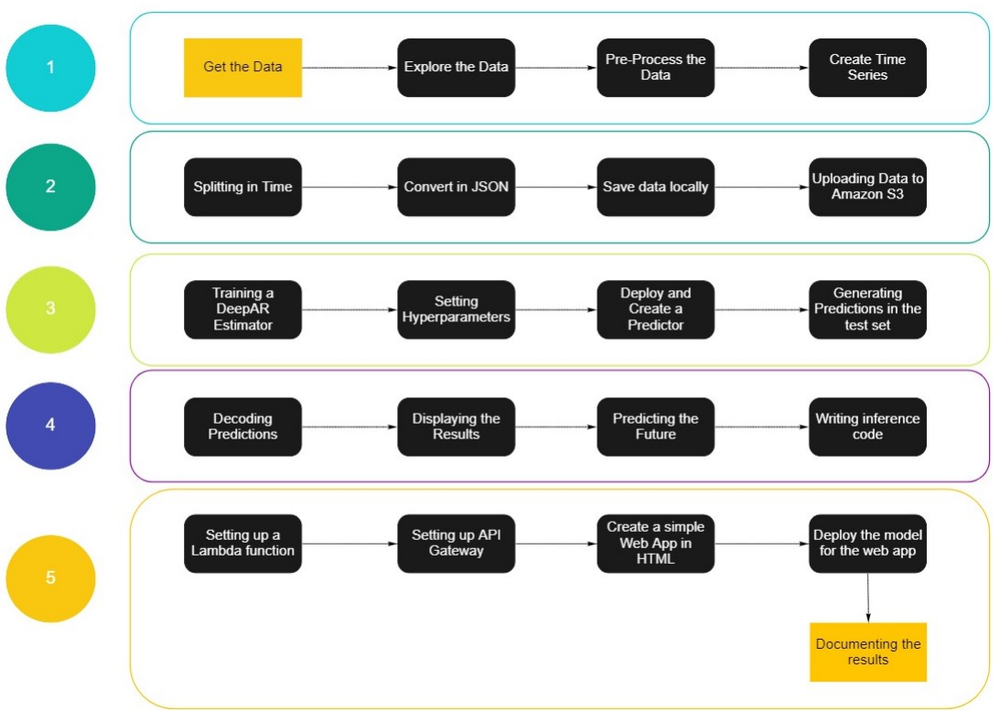

Figure 5: Steps to develop the recommendation system. [4]

The project used as reference to define all development process was the energy consumption prediction during the "4.5. Times Series" class [4].

References

[1] Britannica. Howard schultz, american businessman.

[2] RetailDive. Starbucks rolls out largest mobile payments, loyalty play in us.

[3] Udacity. 3.6. creating a sentiment analysis web app. machine learning in production.

[4] Udacity. 4.5.1. machine learning engineer nanodegree program. time-series forecasting. energy consumption model result.

[5] Udacity. 4.5.1. time-series forecasting. machine learning engineer nanodegree program. time-series forecasting illustration.