Machine Learning Model Assignment

CSCI 59000 – Big Data Analytics

Yuni Xia
yuxia@iu.edu

Shreyansh Mohnot
smohnot@iu.edu

1. Introduction:

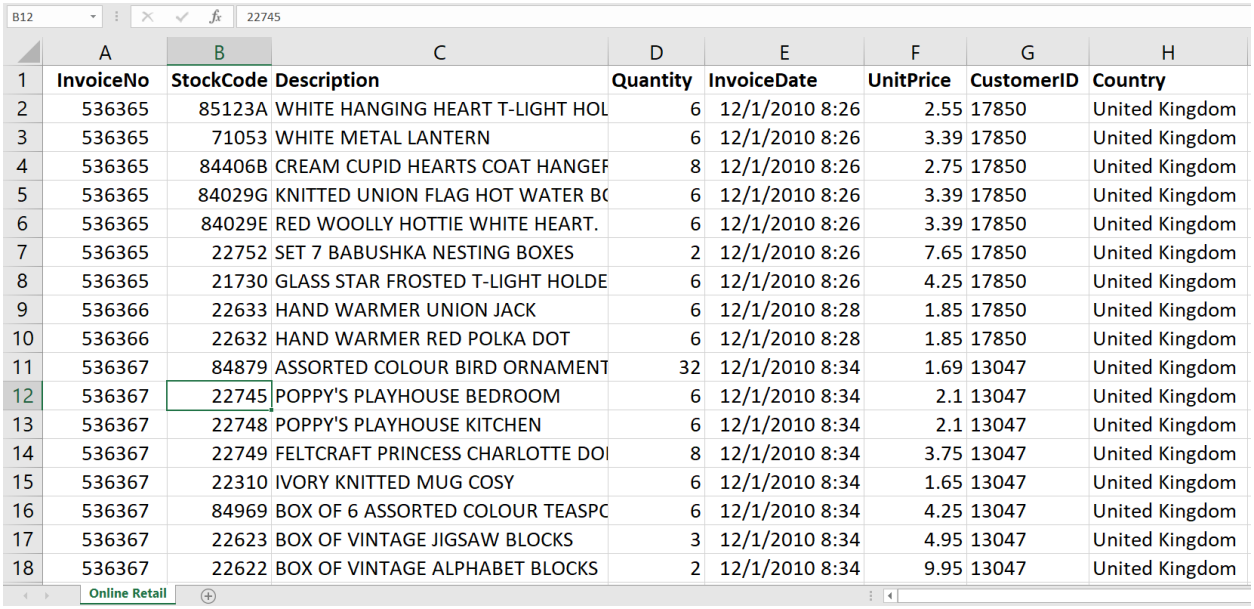
Machine Learning (ML) is a technology which uses previously build datasets to make decisions and generate patterns of how the data is related. It discovers those patterns and constructs models using highly sophisticated machine learning algorithms. These models are then used for making future predictions about the problem.

Machine learning algorithms use training datasets as input for machine learning. It contains all the data points from the past. It is a necessary component in machine learning because each learning process extracts data from training data sets and build machine learning models.

From each training data set a target attribute is selected that is used to predict. It connects those attributed with the data and builds patterns according to it. Thus, whenever the pattern is used is uses the metadata present to give the result for the problem.

2. Data Set:

I have used an Online Retail Database. It is a transactional dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.



B12 | 22745

	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOL	6	12/1/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGEF	8	12/1/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BC	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDE	6	12/1/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOI	8	12/1/2010 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPC	6	12/1/2010 8:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/1/2010 8:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/1/2010 8:34	9.95	13047	United Kingdom

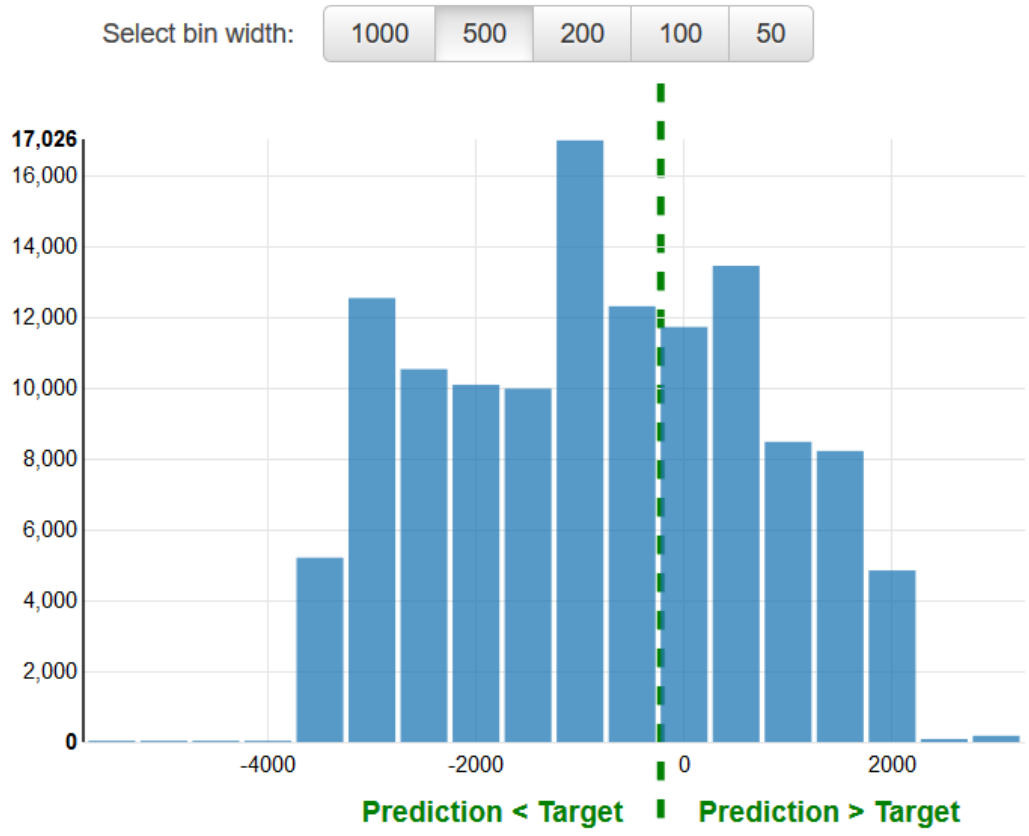
Online Retail

The attributes of the data set are:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

3. Result:

On building Machine learning model on AWS for the data set above, I evaluated the data set on one specific target attribute "CustomerID". Using that Machine learning was able to provide a very satisfactory model for machine learning prediction. The target prediction graph for the data set with target attributes as done in AWS.



ML model performance

On your most recent evaluation, **ev-7DtQaiFi04W** , the ML model's quality score is **better** than the baseline.

RMSE: 1,638.7567
RMSE baseline: 1,711.597
Difference: 72.840

It generated Root-Mean-Square Error RMSE which is used to assess how a system learns a specific model. My finding of RMSE is that it represents the sample standard deviation of the differences between predicted values and observed values.

4. Conclusion:

Machine learning approaches for any data sets involves generating patterns and models for the given target data. However, each pattern requires a target attribute to get the prediction. But it may be not if target value does not reflect the datasets and patterns found are unable to predict the future problems involving the same data. Thus, we can infer that the model may be accurate when the target attribute selected is appropriate in the data sets.

Hence, machine learning on datasets requires highly correlated data which knowledge of data attributes such that patterns predicted, and models generated take even those attributes which are outliers or missing values in the sets.

5. References:

- http://archive.ics.uci.edu/ml/datasets/online+retail
- https://aws.amazon.com/aml/faqs/