

Contents

[**Introduction**](about:blank)

[**Data Processing**](about:blank)

[**Target Variables**](about:blank)

[**Predictive Variables**](about:blank)

[**Pre-modeling**](about:blank)

[**Modeling**](about:blank)

[**Algorithmic Solution Results**](about:blank)

[**Deliverables**](about:blank)

Introduction

* The Online Retail II data set, which gathered actual online retail data in two years, is from the UCI machine learning repository. The source of the data set is Dr. Daqing Chen, Course Director: MSc Data Science. chend **'@'** lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.
* The Online Retail data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers.
* Total sales will be the target variable and predicting variable will be time series.
* Time series analysis model (MA, AR, ARIMA, etc.) and machine learning model will be implemented to analyze and forecast future sales.

The purpose of this document is to provide a detailed technical overview of

1. Data collecting and cleaning
2. Data inclusions and exclusions
3. Predictive variable creation process
4. Target variable definition
5. Iterative model building process
6. Metric evaluation process

Diagram

Description automatically generated

Data Processing

**Data Overview**

We are doing Forecasting on UCI Online Retail Data Set. This dataset contains all the transaction information 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.

**Data Extraction Process**

The data is downloaded from UCI Machine Learning Repository, Online Retail Data.

**Data Diagnostics**

A series of quality checks are performed on the data extract provide. These checks included:

* Attribute information
* Number of records
* Duplicate records if any
* Missing values in relevant fields

Attribution information:

1. 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.
2. StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product (item) name. Nominal.
4. Quantity: The quantities of each product (item) per transaction. Numeric.
5. InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
6. UnitPrice: Unit price. Numeric, Product price per unit in sterling.
7. CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. Country: Country name. Nominal, the name of the country where each customer resides.

Number or records:

1. Year 2009-2010: 525461
2. Year 2010-2011: 541910

Duplicate records if any: There are no duplicate records.

Missing values in relevant fields:

Year 2009-2010

| Attributes | Number of missing values |
| --- | --- |
| Invoice | 0 |
| StockCode | 0 |
| Description | 2928 |
| Quantity | 0 |
| InvoiceDate | 0 |
| Price | 0 |
| Customer ID | 107927 |
| Country | 0 |

Year 2010-2011

| Attributes | Number of missing values |
| --- | --- |
| Invoice | 0 |
| StockCode | 0 |
| Description | 1454 |
| Quantity | 0 |
| InvoiceDate | 0 |
| Price | 0 |
| Customer ID | 135080 |
| Country | 0 |

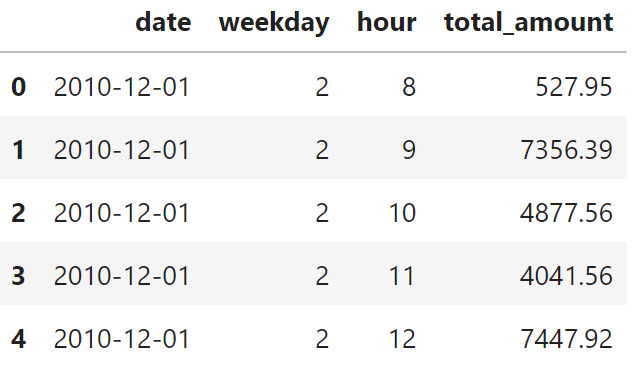
**Modeling Data Creation**

For each of the customers, the modeling data are collected and processed using the following steps. The below also addresses the feature engineering section.

* Holidays: create features using time lags from previous 3 days to future 2 days [-3, 2]
* Time related features: month, year, week are created using datetime

Target Variables

Total\_Amount: quantity times price for each product.



Predictive Variables

A predictive variable is a variable used in algorithmic solutions to predict the target variable. During our analysis, we categorized predictive variables into two categories:

1. Direct variable – These variables were directly from the dataset that was provided by direct customers
2. Derived variable – These variables were created by manipulating the direct variables

**Variable List**

1. Weather: Temperature, Seasons data.
2. Inventory: customer’s inventory.
3. Holiday: US national holidays. Holiday has effect on people’s purchasing patterns, for example, Black Friday, Christmas, etc. It is partly related to the promotion.
4. Features of time: Year, Month, Week. Features of time are critical in the modeling because they capture seasonality in shipments. For example, the shipment of ice cream increases during summer times.
5. How many weeks the product was sold

Pre-modeling

* Before modeling, some analysis techniques were used to discover the seasonality and trend of the original dataset.
* Seasonal decomposition for the hourly sales dataset and daily sales dataset can help us pick the dataset for modeling.
* The pattern and the cut-offs in ACF and PACF can help choose parameters for models such as ARIMA and SARIMAX.

**Training & Testing Split**

Cross validation was used to train the models.

Chart

Description automatically generated

As displayed in the figure below, large portion of the data was used to train the model, and a small portion was used as testing to evaluate the performance of the model. As new data comes in, the model was re-trained and re-tested using the same ratio split. The training and testing ratio splits depend on each run.

Modeling

**Introduction**

After the candidate predictive variables were finalized in the pre-modeling step, the team performed the necessary data mining and statistical iterations to develop and build the Algorithmic Solutions.

1. SARIMAX

* Seasonal Auto-regressive Integrated Moving Average with eXogenous factors, is an extension of the ARIMA class of models.
* SARIMA has the capability of dealing with seasonality and handling exogenous variables.

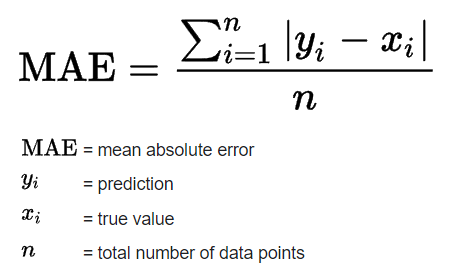
**Hyperparameters Selection of Algorithms**

SARIMAX

Grid Search

This algorithm simply conducts an exhaustive search over all combinations of parameters. The best one will be chosen according to the criteria picked by our team.

**Evaluation Metric**

One metric was adopted in model evaluation: Mean Absolute Error

Mean Absolute Error captures the error between the actual and predicted value which would be a good choice when considering choosing time-series model.

**Algorithmic Solution Finalization**

The following procedure was used to determine the Sales predictions. First, for each product, we run the four models with configured hyperparameters on training data. The model with the highest forecast accuracy is chosen. And then, we train on the whole dataset using the selected model and make predictions of future 20yy.

There is no one-model-fits-all solution because shipment trends differ from customers and product categories. For example, some of the products have a clear seasonality, while some do not. It is also observed that for certain products, the customer decided to change the distribution, which will also add complexity to modeling. In these cases, one model might outperform the others.

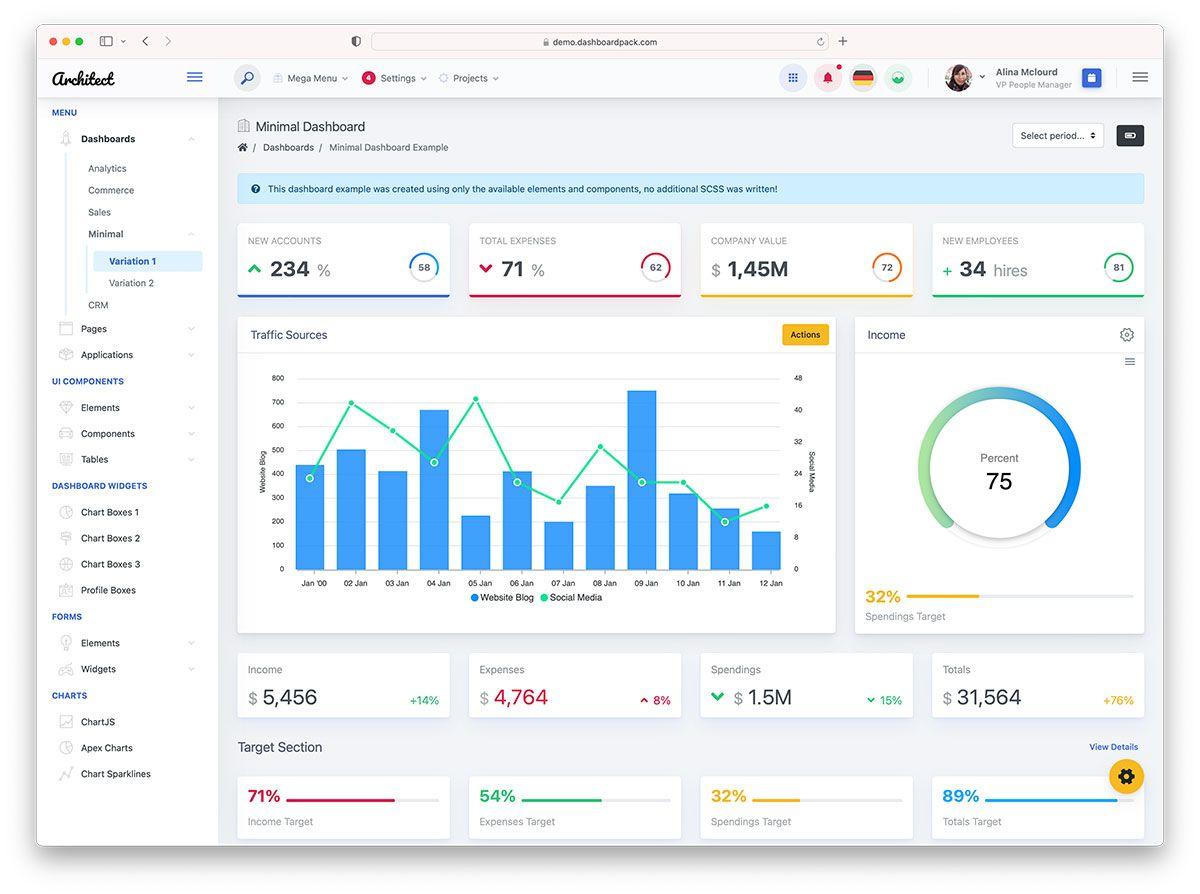
Deliverables

**Variables**

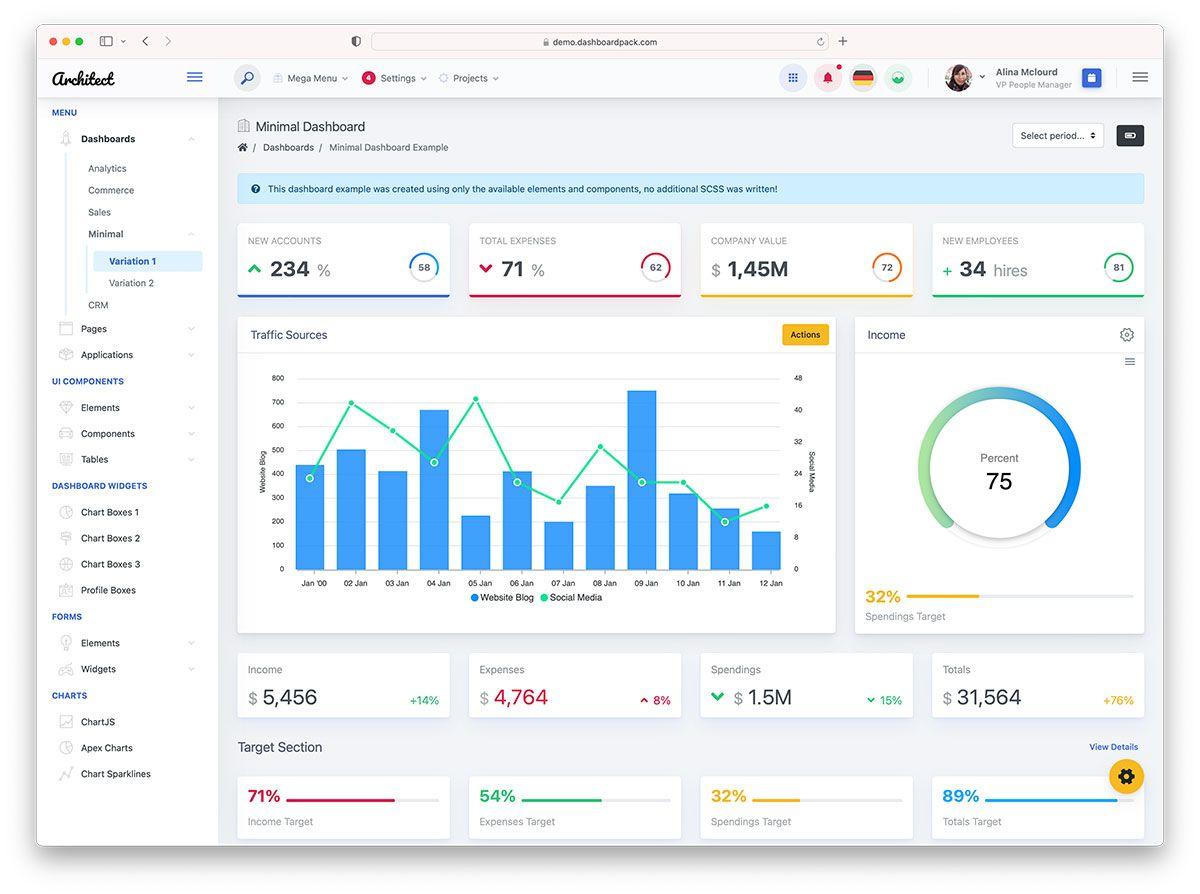
* THF (total head of forecast) – supply chain forecast
* Sales forecast – sss
* Temperature forecast – xxxx
* Electric Units – xxxx

Description of the granularity of the forecast.

**Dashboards**



# Description



# Description

# Team Information