Forecasting Project

Process Documentation

V1.0

Contents

[**Introduction**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840178)

[**Data Extraction**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840179)

[**Data Processing**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840180)

[**Target Variables**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840193)

[**Predictive Variables**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840194)

[**Pre-modeling**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840195)

[**Modeling**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840196)

[**Algorithmic Solution Results**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840197)

[**Deliverables**](file:///E:\Syed\CDML_Documentation.docx#_Toc23840198)

Introduction

* Include different sources of data.
* Apply advanced analytics to develop a robust approach to develop ML model for North America in order to maximize growth and turnover.
* Standardize a robust methodology (including analytical approach and data requirements) for the development of forecasting that will be leveraged in coming years.
* Customers include: List of customers.

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

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

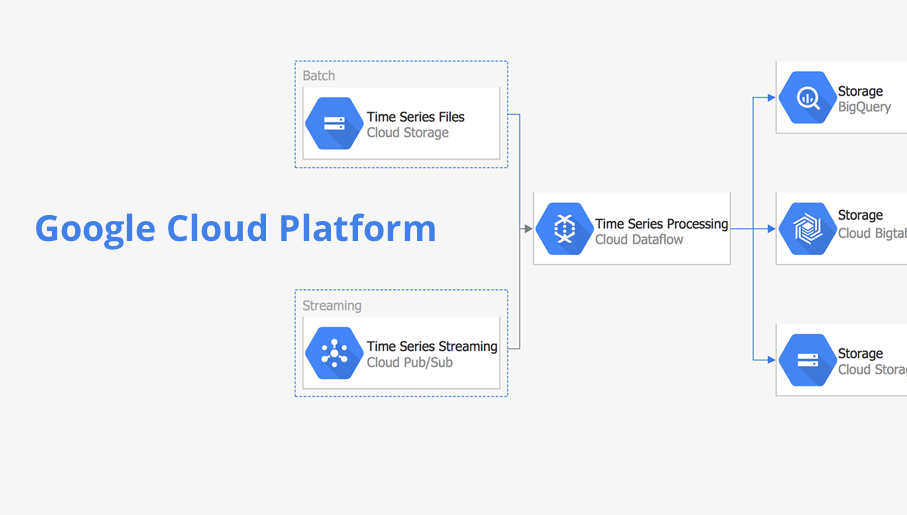
Diagram

Description automatically generated

Data Extraction

**Workflow**

The section explains about workflow of how the data was extracted and stored. Data is received in different formats, and the below workflow is used to extract, process and store the data.

 For customer Target, internal and external data sources are uploaded to One Drive, and all processing, modeling and reporting are done in GCP.

For other customers, internal and external data sources are collected and stored locally. Data are processed and integrated using Python scripts. The processed files are then uploaded to One Drive.

Data Processing

**Data Overview**

The purpose of this section is to detail the data extraction process, data exclusions and transformations, and the diagnostics performed on data received.

All data extracted is weekly data from 20xx, and forecasting period is till the end of 20yy.

**Data Extraction Process**

Data extraction is based on each customer: List of customers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Customer | Data | Source | Internal/External | Format | Where it is stored |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Data Diagnostics**

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

* Number of records
* Duplicate records if any
* Missing values in relevant fields
* Sum of a numeric field
* Data period confirmation

Diagram

Description automatically generated

**Modeling Data Creation**

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

A target variable is the variable to predict as a result in algorithmic solutions.

Temperature in Fahrenheit, Sales in dollars, shipments, etc. could be the target variables.

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

* The forecast is done till the end of this year (20yy). For holidays and promotions of the future, the data is acquired regularly. For external projections, and the projections are acquired.
* The timeline of data collected is from 20xx till date. For those which were produced after 20yy, the dates before their launch were dropped. In this way, we don’t create noise in modeling during the pre-launch period.
* Due to COVID and current situation, assumptions made to accommodate in the model list here. For example, a variable indicate change of policy is incorporated in the model

**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**

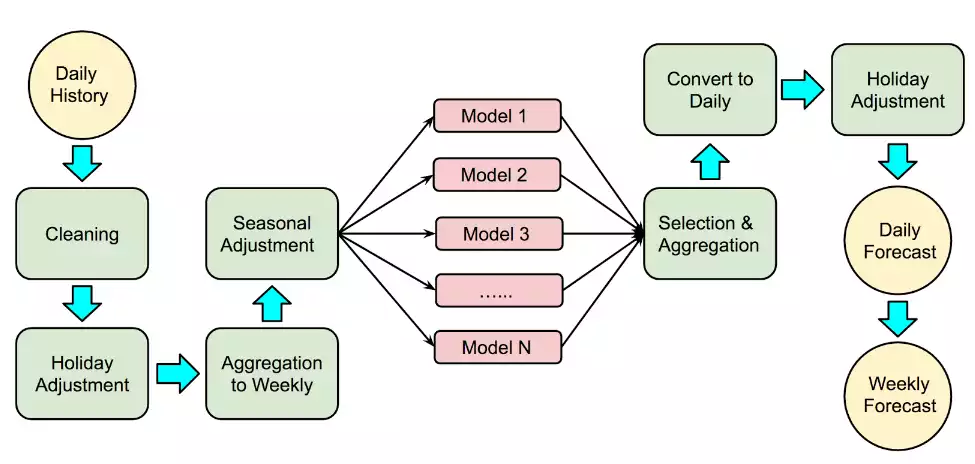
After the candidate predictive variables were finalized in the pre-modeling step, the team performed the necessary data mining and statistical iterations to develop and build the Algorithmic Solutions. There are a variety of ways in which Algorithmic Solutions can be developed including linear regression, decision trees, neural networks, etc. In this engagement, four models were utilized in the development of the Algorithmic Solutions: Random Forest, Gradient Boosting, LASSO Regression and Neural Networks. During each run, we run the four models and the one with the best performance is adopted. In the following section, we’ll illustrate how model is built, what evaluation metric was used and the comparisons between different models.

The Algorithmic Solution was built for three major targets and Others; see the chart below for categories.

**Algorithmic Solution Design**

|  |  |  |
| --- | --- | --- |
| **Target** | **Category** | **Algorithms/Approach** |
|  |  | Random Forest  Gradient Boosting  LASSO  Neural Net |
|  |  | Random Forest  Gradient Boosting  LASSO  Neural Net |
|  |  | Random Forest  Gradient Boosting  LASSO  Neural Net |
|  |  | Random Forest  Gradient Boosting  LASSO  Neural Net |

Four models were trained for each product individually. For each of the Algorithmic Solutions built, the following modeling specifications were applied:



1. Random Forest

* Used random forest method
* The number of trees in the forest (n\_estimators)
* The number of features to consider when looking for the best split (max\_features)="sqrt"
* The min number samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min\_samples\_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model
* The minimum number of samples required to split an internal node (min\_samples\_split)
* A random number variable was also generated for variable selection in the random forest model. All the variables behaving better than this random variable were considered in each of the model iteration

1. Gradient Boosting

* Used Gradient Boosting tree-based method, which is like Random Forest
* The number of trees in the forest (n\_estimators)
* The number of features to consider when looking for the best split (max\_features)="sqrt"
* The min number samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min\_samples\_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model
* The minimum number of samples required to split an internal node (min\_samples\_split)
* A random number variable was also generated for variable selection in the random forest model. All the variables behaving better than this random variable were considered in each of the model iterations

1. LASSO Linear Regression

* Used LASSO Regression method, which is a linear function
* Alpha – the punishment weight of LASSO Regression; it is used to smooth the model and reduce variance
* Max\_iter – the maximum number of iterations, this is used to avoid overfitting and long computation time

1. Neural Net

* Used Neural Network, which is non-linear
* Learning rate controls how much to change the model in response to the estimated error each time the model weights are updated
* Batch size is the number of training examples utilized in one iteration. It controls the accuracy of the estimated error when training the neural network.
* Number of epochs refer to the number of times the learning algorithm runs through the entire training dataset. It controls the error rates and training time.

**Comparisons of Algorithmic Solutions**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Pros** | **Cons** |
| Random Forest | Easy to interpret  Does not tend to overfit | Lose some correlation between features |
| Gradient Boosting | Better capture non-linear relationships | Overfitting |
| LASSO Regression | Good for clear trends | Does not work for unstable situations |
| Neural Net | Good for complex trends or change is not constant | Interpretability  Overfitting |

**Hyperparameters Selection of Algorithms**

For each of the four models. there are two set of hyperparameters: one used in normal situations when sales and shipments follow the trends. Another was used in abnormal situations, particularly during COVID time when sales are highly unstable, and trends are more complicated.

The two set of hyperparameters are pre-configured and can be applied depending on current situations.

**Evaluation Metric**

Two metrics were adopted in model evaluation: forecast accuracy and forecast bias.

Text

Description automatically generated

Forecast accuracy measures the absolute difference between the forecast vs the reality, while forecast bias measures how far apart is the forecast from actual outcomes.

When selecting the best models, the one with the highest forecast accuracy is adopted. Forecast bias is included in reporting and dashboards for references.

**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.

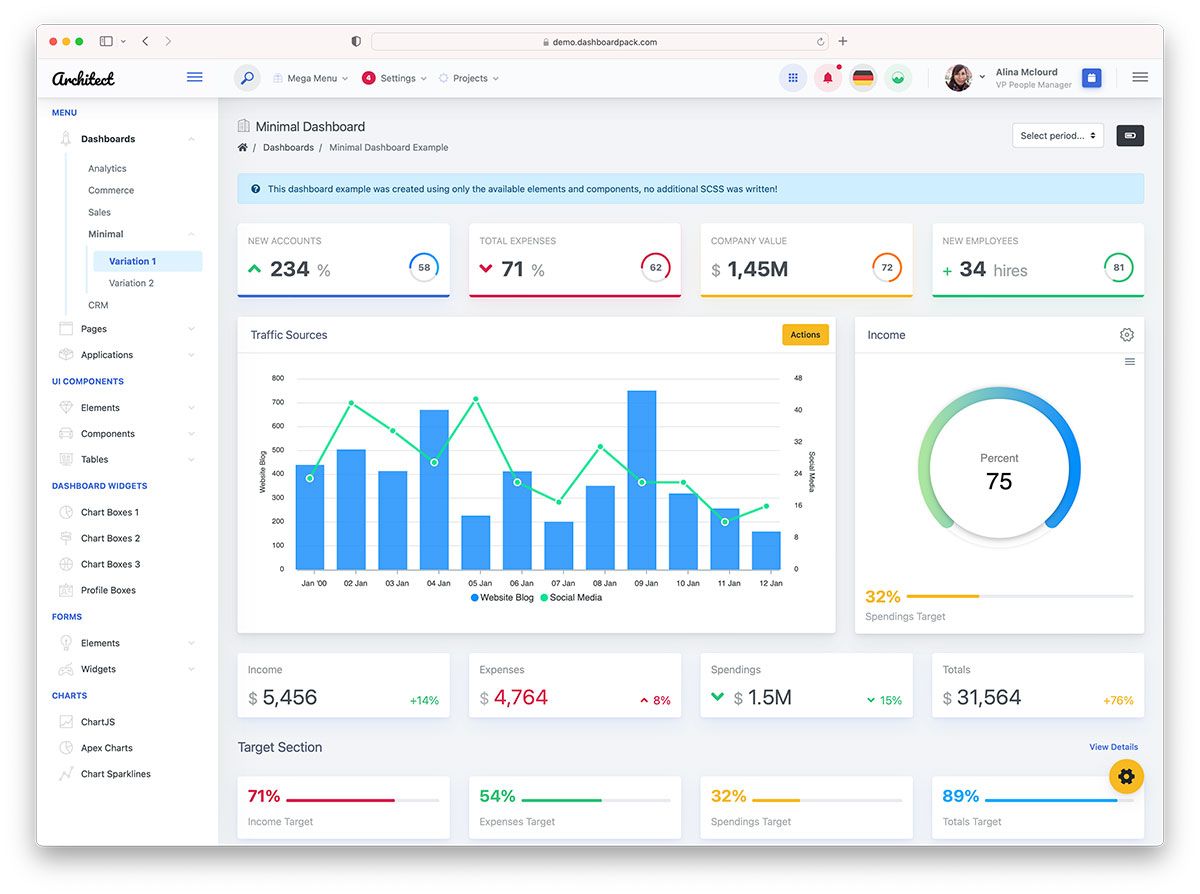
Deliverables

**Variables**

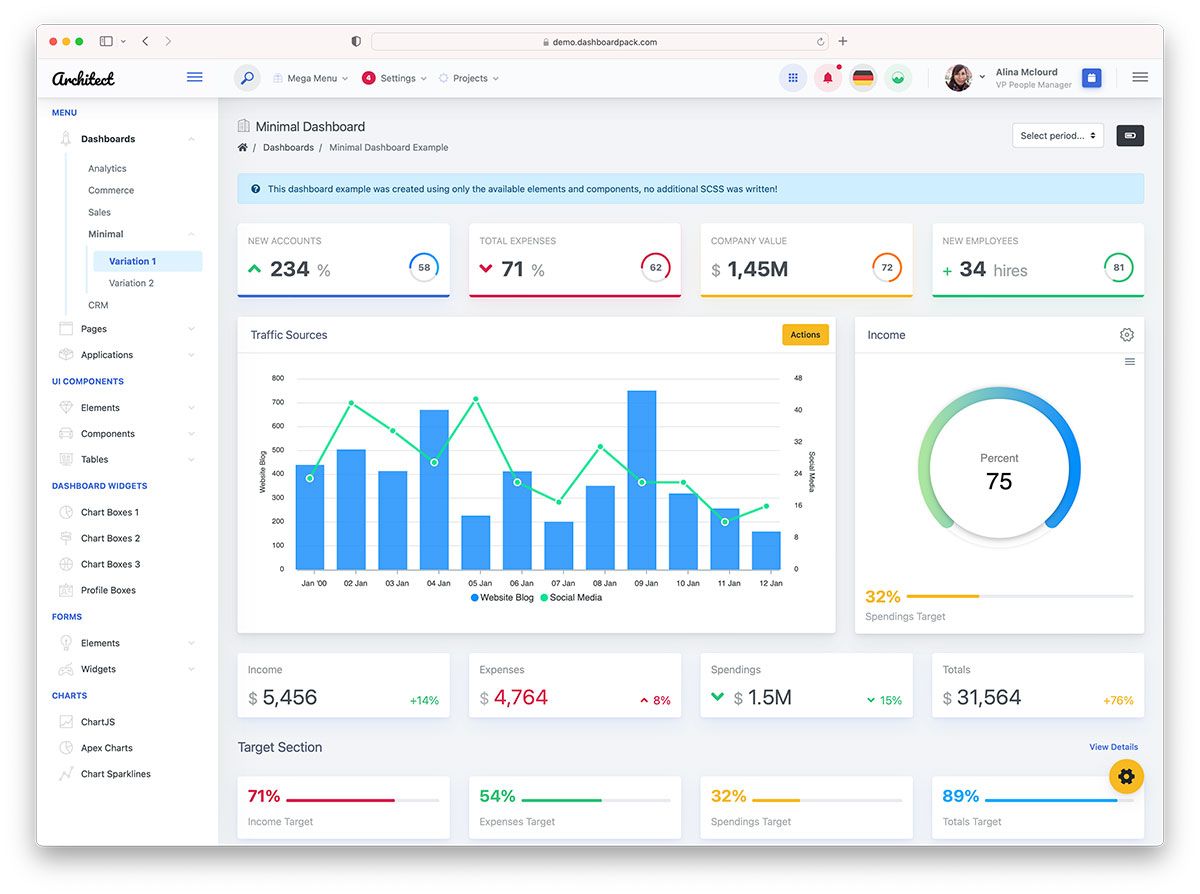
* THF (total head of forecast) – supply chain forecast
* Sales forecast – sss
* Temperature forecast – xxxx
* Electric Units – xxxx

Description of the granularity of the forecast.

**Dashboards**



# Description



# Description

# Team Information