Template found here:

[pair-code.github.io/datacardsplaybook](http://pair-code.github.io/datacardsplaybook)

| **Store Sales - Time Series Forecasting** | Intended Use:  The store sales dataset offers a comprehensive collection of time-series data pertaining to store sales in Ecuador. It encompasses information such as store location, product details, and daily sales records over a period of time. It is a good tool to test time series models on. |
| --- | --- |
| DATASET LINK | DATA CARD AUTHOR(S) |
| https://www.kaggle.com/competitions/store-sales-time-series-forecasting | **Alec Bothwell**  **William Earley**  **Cory LeRoy** |

| Authorship Publishers | | |
| --- | --- | --- |
| **Publishing Organization**Kaggle.com | **Industry Type**  Academic  Tech | **Citation**  Alexis Cook, DanB, inversion, Ryan Holbrook. (2021). Store Sales - Time Series Forecasting. Kaggle. https://kaggle.com/competitions/store-sales-time-series-forecasting |
|  |  | |

###### 

###### 

| Motivations & Intentions | | |
| --- | --- | --- |
| Motivations | | |
| PURPOSE(S) | DOMAIN(S) OF APPLICATION | MOTIVATING FACTOR(S) |
| Research, Education | *Machine Learning, Education, Supply Chain* | This project aims to investigate the efficacy of time-series forecasting ML models in mitigating food waste within grocery stores. The imperative of curbing food waste, while ensuring adequate sustenance for individuals, underscores its profound implications for environmental sustainability. |

| Dataset Overview | | |
| --- | --- | --- |
| DATA SUBJECT(S) | DATASET SNAPSHOT | CONTENT DESCRIPTION |
| train.csv | | Size of Dataset | 118,946 KB | | --- | --- | | Number of Instances | 3,000,888 | | Number of Fields | 6 | | **Interpret a datapoint:**  Each datapoint is the summed value of sales for a family in store on a given date. This is to be joined to other supplementary datasets given and then transformed during feature extraction |
| **Summary of data:**   | Field | id | date | store\_nbr | family | sales | onPromotion | | --- | --- | --- | --- | --- | --- | --- | | Summary | 3,000,888 distinct values | 2012-12-31 to 2017-098-14 | 54 distinct values | 33 distinct values | See summary statistics below | See summary statistics below | | Example value | 41 | 2013-01-01 | 1 | BEVERAGE | 172.0 | 0 | | | |
|
|
|  |  |  |
| **Summary Statistics:**  | Statistic | date | onpromotion | sales | | --- | --- | --- | --- | | count | 2,765,697 | 2,765,697 | 2,765,697 | | mean | 2015-05-21 | 2.84 | 363.32 | | std |  | 12.7 | 1,074.40 | | min | 2013-01-01 | 0 | 0 | | 25% | NA | 0 | 0 | | 50% | NA | 0 | 14.00 | | 75% | NA | 0 | 217.75 | | max | 2017-08-15 | 741 | 124,717.00 | | | |
| DESCRIPTIVE STATISTICS | | |
| holidays.csv | | Size of Dataset | 23 KB | | --- | --- | | Number of Instances | 350 | | Number of Fields | 6 | | **Interpret a datapoint:**  Each datapoint is a holiday. The locale\_name field indicates if the holiday is at the city, state, or national level. Transferred indicates if the holiday is observed/celebrated on a different day |
| **Summary of data:**   | Field | date | type | locale | local\_name | description | transferred | | --- | --- | --- | --- | --- | --- | --- | | Summary | *2012-03-01 to 2017-12-25* | *Holiday 63%*  *Event 16%*  *Other (73) 21%* | *National 50%*  *Local 43%*  *Regional 7%* | *24 distinct value* | *103 distinct values* | *True 3%*  *False 97%* | | Example Value | *2014-11-03* | *Holiday* | *National* | *Ecuador* | *Independencia de Cuenca* | *False* | | | |
| DESCRIPTIVE STATISTICS | | |
| stores.csv | | Size of Dataset | 2 KB | | --- | --- | | Number of Instances | 54 | | Number of Fields | 5 | | **Interpret a datapoint:**  Each datapoint is the store\_nbr ID with the location and cluster it belongs to |
| **Summary of data:**   | Field | store\_nbr | city | state | type | cluster | | --- | --- | --- | --- | --- | --- | | Summary | 54 unique values | 22 unique values | 16 unique values | 5 unique values | 17 unique values | | Example Value | 2 | Quito | Pichincha | D | 13 | | | |
| df\_oil | | Size of Dataset | 22 KB | | --- | --- | | Number of Instances | 1175 | | Number of Fields | 2 | | **Interpret a datapoint:**  Each datapoint is the price of oil on a given date |
| **Summary of data:**   | Field | date | Dcoilwtico (oil price) | | --- | --- | --- | | Summary | 2012-12-31 to 2017-08-30 | See summary statistics below | | Example Value | 2013-01-02 | 93.14 | | | |
| **Summary Statistics:**  | Statistic | oil\_price | | --- | --- | | count | 1,175 | | mean | 67.71 | | std | 25.63 | | min | 26.19 | | 25% | 46.41 | | 50% | 53.19 | | 75% | 95.66 | | max | 110.62 | | | |
| DATA MERGING | | |
| Data is merged as shown below | | |

| Example of Data Points Post Joins |
| --- |
| | **Field Name** | **Field Value** | **Description** | | --- | --- | --- | | date | 2013-01-11 | Datetime64 of sales date | | store\_nbr | 1 | Int64 of id for the store | | family | BEVERAGE | Object value of the category of item sold | | onPromotion | 0 | Int64 value of number of items on promotion | | storeCluster | 13 | Int64 value which represents the ID of the cluster a store belongs to | | holiday | Carnival | Object value of the name of the holiday on a given day | | oil\_price | 97.01 | Float64 value of the price of oil on a given day | | sales | 172.0 | Float64 value of total sales for the product category in a store on the given date | |

| Transformations | | |
| --- | --- | --- |
| Fill this section if any transformations were applied in the creation of your dataset. | | |
| Synopsis | | |
| TRANSFORMATION(S) APPLIED | FIELD(S) TRANSFORMED |  |
| Remove ValuesConverting Data TypesCleaning Missing ValuesDimensionality ReductionScaled Values | **<Transformation Type>**   | **Field Name** | **Source & Target** | | --- | --- | | date | Remove values | | Converting data type | | oil\_price | Scaled values | | Cleaning missing values | | onPromotion | Scaled values | | family | Dimensionality reduction | | holiday | Dimensionality reduction | | |
| Breakdown of Feature Engineering and Transformations | | |
| CLEANING MISSING VALUE(S) | METHOD(S) USED | COMPARATIVE SUMMARY |
| **Oil\_price** | **Summary:** Oil price has missing values where there is not a date or price available on certain days. And null values when there is a date but a null oil price.  **Method:** Missing oil prices were interpolated as the mean of the values before and after the missing oil price. | 524 dates have oil price interpolated |
| REMOVE VALUES |  | COMPARATIVE SUMMARY |
| **Date** | **Summary:** Every date has a sales value for each store and family. This includes dates when the store is closed. This will unintentionally bring down the forecast.**Method:** Extended (month long) periods of 0 sales are removed from the dataset | Initial dataset is 3,000,888 rows. After 0 sales dates are removed, dataset has 2,765,697 rows |
| CONVERTING DATA TYPES | METHOD(S) USED | COMPARATIVE SUMMARY |
| **Date** | Date was transformed into day\_of\_week, month, year | New date features   | **Field Name** | **type** | | --- | --- | | day\_of\_week | int | | month | int | | year | int | |
| DIMENSIONALITY REDUCTION | METHOD(S) USED | COMPARATIVE SUMMARY |
| **family** | **Summary:** Family has 33 unique categories. During data exploration, we noticed certain families had similar sales trends. Dimensionality reduction is used to allow models to generalize better **Method:** K-means clustering used to group the families into clusters | 33 unique categories reduced to 6 |
| **holiday** | **Summary:** The dataset gives 103 unique holidays. Many are minor holidays that don’t effect sales. However, a few also have a huge impact on the daily sales both negatively and positively **Method:** One hot encode the holidays and then run Ordinary Least Squares to see which holidays are statistically significant. Remove any that are not | Of 103 unique holidays, 83 are dropped leaving 18 remaining |
| SCALED VALUES | METHOD(S) USED | ADDITIONAL CONSIDERATIONS |
| Models are impacted when the range of values varies between different features. Values need to be normalized so no one feature can dominate the model | onPromotion, oil\_price were scaled to reduce range of values using MinMaxScalar |  |

| Use in ML or AI Systems | |
| --- | --- |
| DISTRIBUTION(S) | ADDITIONAL NOTES |
| Train 1,935,987  Validation 276,569  Test 553,141  Train, Validation, Test split. 70%-10%-20%. | Additional feature extraction may be done depending on what the model needs. Ex. creating dummy variables for categorical features. |

###### 

###### 