| **Model Card Version:** 0.0\_2024 Ordinary Least Squares (OLS) **Model Card Authors:** William Earley, Alec Bothwell, Cory LeRoy | OLS is a simple linear regression model which minimizes the sum of residuals between the observed and predicted values. OLS may not have the best predictive power, but it is easy to interpret. This is why it will be used as a baseline model. It is easy to see the impacts of each feature on the prediction due to The model outputs coefficients of the features from the model which makes it easy to see the impact and relationship between each feature compared to the value being predicted. |
| --- | --- |
|  | |

| Model Snapshot | | | |
| --- | --- | --- | --- |
| Model Overview | | | |
| MODEL ARCHITECTURE | INPUT(S) | | OUTPUT(S) |
| The model is summarized by the following equation:  Y is dependent variable being predicted which is sales  are the features  are the coefficients of the features which quantify the impact of each feature  is the error term  OLS finds the values of which minimizes the following optimization function:  Where Yi is the actual value and is the predicted value | Dataframe of features. All features must be int of float values. Categorical features must be made into dummy variables with values either 0 or 1 | | Float value of sales prediction |
| Usage | | | |
| APPLICATION | BENEFITS | | KNOWN CAVEATS |
| Where has this model been used, or where is it currently used? Include links for readers to learn more. | Why might users choose to use this model, relative to others? Evidence your response with metrics or performance results | | Are there any known and preventable failures about this model? |
| OLS is an econometrics model that has been used in every industry.  Statsmodels was installed using:  Pip install statsmodels | OLS is highly interpretable. It provides coefficients which quantify the impact of each feature on the target variable. It also gives a p value for each feature which if less than a significance value (usually .05) can indicate if the feature has a statistically significant effect on the model's predictions. | | Not the most powerful predictor, especially for time series data. Other models have time series built in with additional inputs for seasonality as well. OLS has none of this. |
| Model Creators | | | |
| CITATION Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*. | | | |
|
|
| System Type | | | |
| SYSTEM DESCRIPTION | UPSTREAM DEPENDENCIES | | DOWNSTREAM DEPENDENCIES |
| The predictions of OLS are to be used as a standalone model. OLS is also used to determine the significance of holidays | The data fed into OLS must be numerical. Categorical fields are widened by converting into dummy variables with values either 0 or 1. OLS also contains a constant term which has to be added.  Model imported using: import statsmodels.api as sm  Model can be created and fit using the below code:  Model = sm.OLS(y,X).fit() | | OLS is used to measure statistical significance of holidays. Not significant holidays are removed from the dataset before being fed into the other models |

| Data Overview | | | | | |
| --- | --- | --- | --- | --- | --- |
| TRAINING DATASET SNAPSHOT | | DATASET MAINTENANCE & VERSIONS | | EVALUATION DATA | |
| Statistics of df\_train.pkl | | Data is preprocessing in the Feature Engineering notebook. Files are df\_train.pkl, df\_val.pkl, and df\_test.pkl | | Train 1,935,987  Validation 276,569  Test 553,141  **Above:** Train, Validation, Test split. 70%-10%-20%. | |
| **Dataset Size** | 184,015 KB |
| **Number of Instances** | 1,935,987 |
| **Number of Fields** | 29 |
|  | |
|  | | | | | |

###### 