| **Model Card Version:** 0.0\_2024 eXtreme Gradient Boosting XGBoost **Model Card Authors:** Cory LeRoy, Alex Bothwell, Will Earley | XGBoost is an ensemble learning algorithm that builds on gradient boosting to perform both regression and classification tasks. XGBoost does a great job handling complex, structured data with high dimensionality and nonlinearity which we have here. The robustness to overfitting, high performance, interpretability and flexibility offered by XGBoost make it a great candidate for our grocery sales problem. |
| --- | --- |
|  | |

| Model Snapshot | | | |
| --- | --- | --- | --- |
| Model Overview | | | |
| MODEL ARCHITECTURE | INPUT(S) | | OUTPUT(S) |
| XGBoost revolves around an ensemble learning technique called gradient boosting. It is made up of multiple decision trees, or weak learners, which are then combined to form a strong learner. Working with decision trees, boosting, gradient descent, regularization, and parallelization, XGBoost combines the strengths of decision trees with advanced boosting and regularization techniques to create a versatile and powerful ensemble learning model. [Full documentation here.](https://xgboost.readthedocs.io/en/stable/) | 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. In this case we had the instances id, store number, onpromotion, cluster, familycluster, day\_of\_week, month, year, oil price, holiday observed, and family dummy variable | | Float value of sales prediction |
| Usage | | | |
| APPLICATION | BENEFITS | | KNOWN CAVEATS |
| Due to its robustness and high performance, XGBoost has a wide range of applications including predictive analysis, recommender systems, natural language processing, image recognition and many more. | Users would choose XGBoost due to its accurate results, scalability, robustness to overfitting, flexibility, interpretability and proven track record of success. In many [kaggle competitions](https://www.kaggle.com/code/dansbecker/xgboost) XGBoost has been used by the top performing teams and has achieved leaderboard dominance. | | There are some limitations to the XGBoost model. If tuned incorrectly, it can suffer from overfitting, it struggles with imbalanced data sets, can be sensitive to hyperparameter settings, and can be computationally intensive. |
| Model Creators | | | |
| Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785 | | | |
|
|
| System Type | | | |
| SYSTEM DESCRIPTION | UPSTREAM DEPENDENCIES | | DOWNSTREAM DEPENDENCIES |
| This is a standalone model designed to be used independently. | N/A | | N/A |

| Data Overview | | | | | |
| --- | --- | --- | --- | --- | --- |
| DATA PRE-PROCESSING | | DEMOGRAPHIC GROUPS | | EVALUATION DATA | |
| In the data pre-processing stage, missing values should be handled, specific features should be engineered, outlier detection and removal, and categorical variables should be handled. | | Not applicable here. | | Nothing notable, test set matches the training set, it represents the following months in the time series after the training set. | |
|  | | | | | |