Predicting Sales for Rossmann Drug Stores

CS 229 | Machine Learning

Brian Knott, Andrew Simpson, and Hanbin Liu

Objective

Rossmann Drug Stores is hosting a Data Science Competition on Kaggle with the goal of predicting sales based on several store features.

Features:

Date Day of Week Store Age
Store Type Promotions
Past Sales Etc.

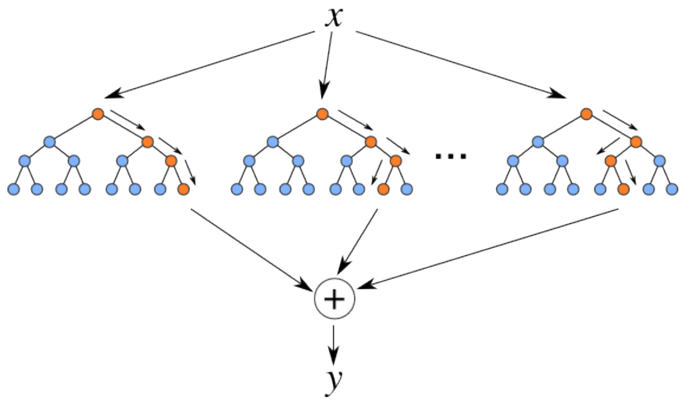
Predict Sales:

Min. Residual Mean Square % Error:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

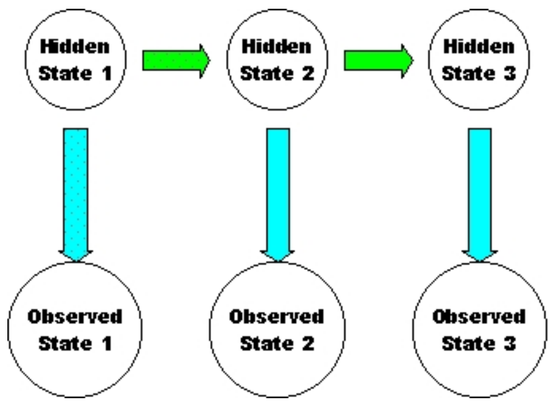
Model #1: Random Forest Regression

Random Forest Regression is an ensemble learning algorithm that operates by aggregating many random decision trees to make predictions while avoiding overfitting. We started by using the Random Forest algorithm because its bagging techniques are robust to data abnormalities and because random forest packages are widely available



Model #2: Hidden Markov Model

Hidden Markov Modeling is a sequence / state estimation model that assumes that the dataset derives from a Markov Process with hidden state information. Hidden Markov Models are appropriate for this dataset since its data evolves over time, a phenomenon unaccounted for in ensemble learning.

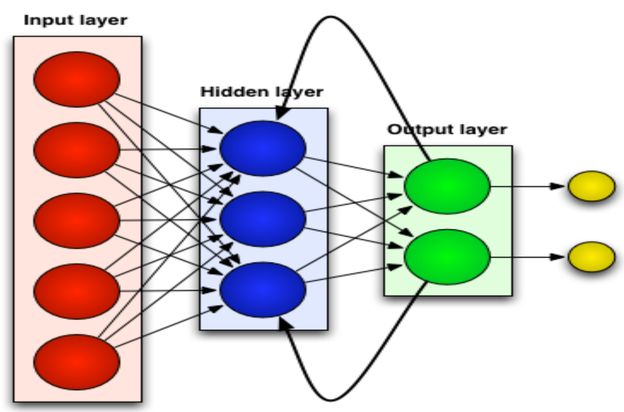


Results

Model	RMSPE
RF	11.98%
HMM	16.52%
RNN	13.20%
GB	11.26%

Model #3: Recurrent Neural Network Regression

Recurrent Neural Networks work similarly to Neural Networks with the addition of feedback loops. The internal recursive states allow this kind of neural network to exhibit dynamic temporal behavior, making it appropriate for processing time series data.



Model #4: Gradient Boosting Regression

Gradient Boosting is an ensemble learning algorithm that uses a weighted average of simple models to learn a more complex model. The algorithm first uses a simple model to fit the data, then a simple model to fit the residuals between the data and the first model. This process continues on each model's residuals until an appropriate fit is found.