Final Project

Sales Prediction

Table of Contents

Team Background

- Competition Name: Predict Future Sales
- Participant Name: Shyju Kozhisseri
- Private Leaderboard Score: 0.926480
- Private Leaderboard Place: 2261
- Location: Canada
- Email: shyjukozhisseri@gmail.com

Background

I’m an ETL developer with a computer science & engineering degree. This is my first competition in Kaggler. It was a part of the Coursera course: How to Win a Data Science Competition. It took me around 3 days to complete the assignment.

Summary

The training model used for prediction was the Random Forest Regressor. First the raw data is converted to feature matrix with 4 months lag. Then it is mean encoded and fed in to the model. Ensembling techniques were not used as these were in fact reducing the accuracy score of the overall model. It takes around 1200 seconds to fully train the model.

Features Selection / Engineering

Simple mean encoding technique is used to encode the categorical features. Only minimal features are selected in order to avoid overfitting.

Training Method

Random Forest Regressor is used for modeling. Ensembling is not used as it was not contributing significantly for the improvement of the score. The hyper parameters are tuned using the functions from the library sklearn model selection.

Interesting Findings

Only 43% of the shop-item combination is present in the test set. Using this knowledge has reduced the effort of computation in training the model.

Model Execution Time

The model training along with feature engineering took around 1200 seconds to complete. Predictions are made in less than 10 seconds.

Tools

numpy 1.18.1

pandas 1.0.3

sklearn 0.22.2.post1

scipy 1.4.1

seaborn 0.10.0

References

Kaggle Final Project: Predict Future Sales