Final Project – Sales Prediction

Coursera: How to Win a Data Science Competition: Learn from Top Kagglers

Name: Rodrigo de Lima Oliveira
Location: Brazil / Santa Catarina
E-mail: rodrigolima82@gmail.com
Linkedin: https://www.linkedin.com/in/rodrigolima82/

Backgroud

Rodrigo Lima Oliveira

- Bachelor’s degree in IT
- Post Graduate in Project Management
- Data Scientist graduate in Data Science Academy

Summary

I used 3 distinctively different models:

- Ridge Regression (Linear)
- LightGBM (Tree based)
- XGBoost (Tree based)

After each model has been trained, a stacked ensemble model which the three models

The most importante features are the lagged month intervals

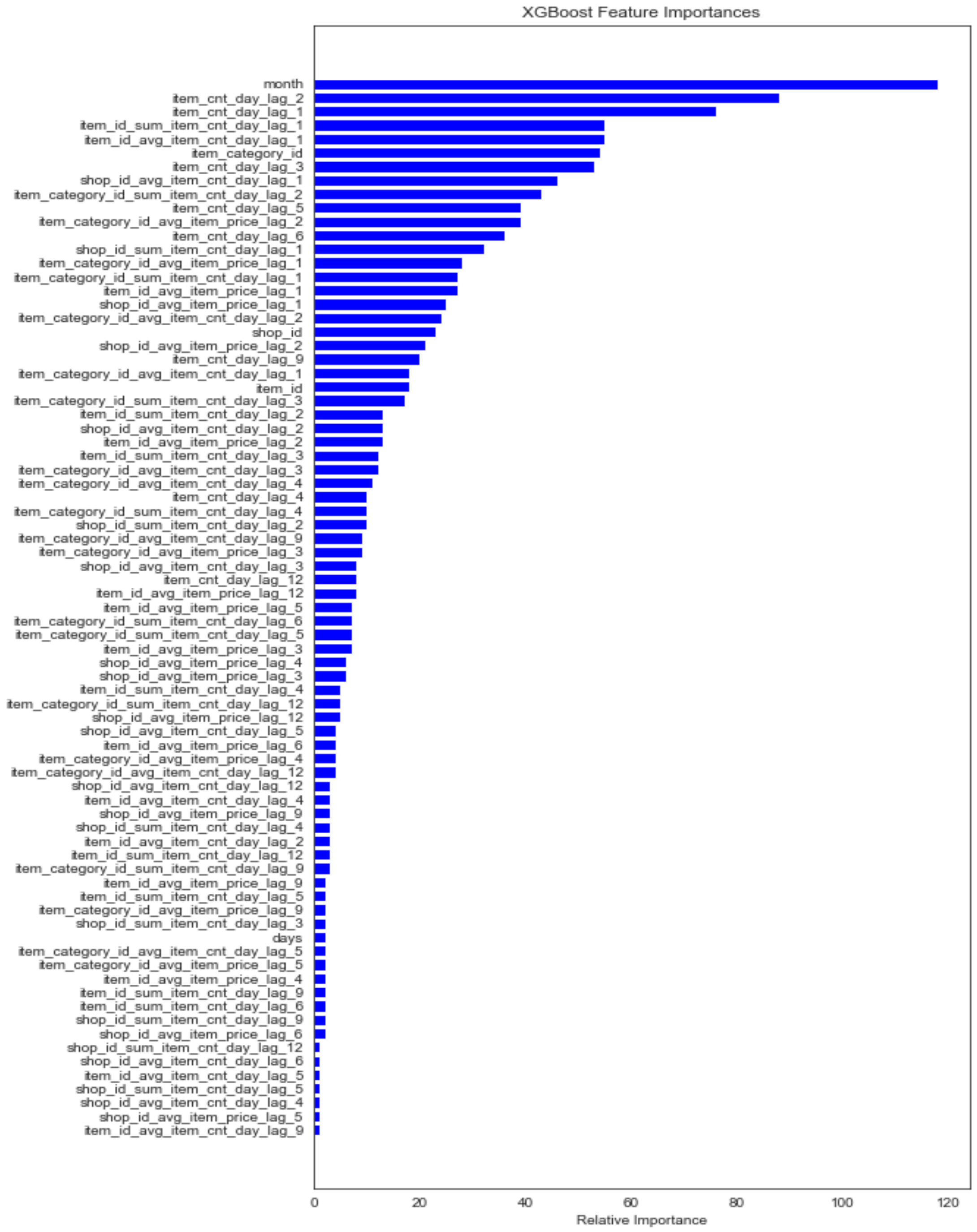
Features Selection / Engineering

MOST IMPORTANT FEATURES ARE THE LAGGED MEAN ENCODED VALUES FROM THE CATEGORICAL DATA

I SELECT FEATURES WITH XGBOOST PLOT ON FEATURE IMPORTANCE

I DID NOT USE ANY EXTERNAL DATA

Most Important Features



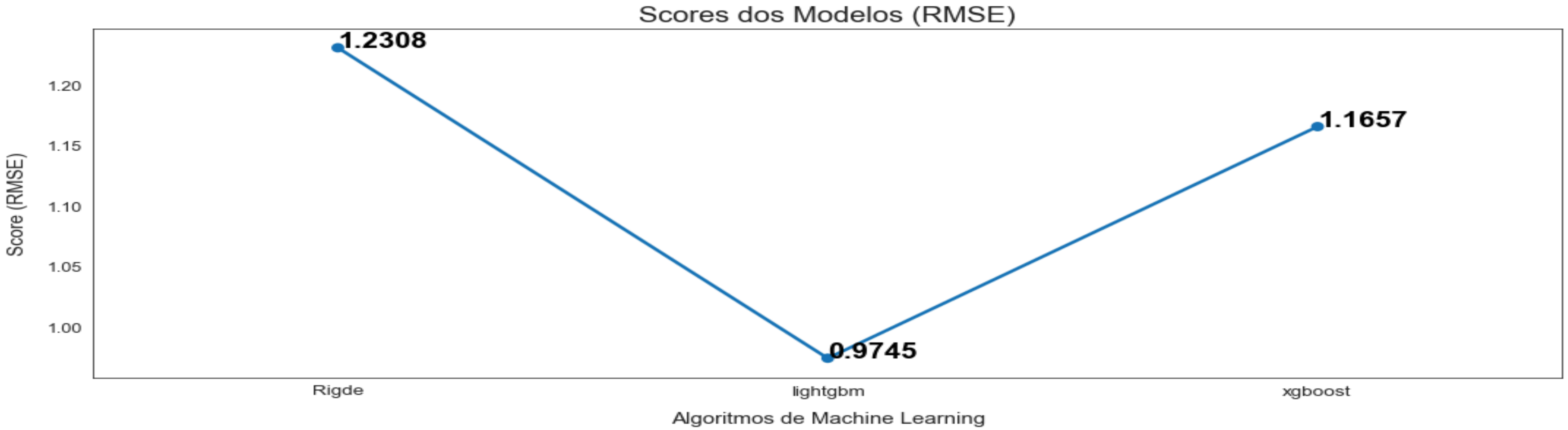
Training Methods

I used 3 distinctively different models:

- Ridge Regression (Linear)
- LightGBM (Tree based)
- XGBoost (Tree based)

I tried on stacking with Random Forest, Extra Tree and Linear Regression as first level model and use XGBoost as second level model without fine tuning and it give terrible result. Due to time constraint, we decide not to use ensemble here

Results



RMSE	Train	Test
Ridge Regression	1.2308	1.2481
LightGBM	0.9745	1.1517
XGBoost	1.1657	1.2088