



ROSSMANN Store Sales Challenge

Time Series Analysis for Sales Forecast

By BYF Royale - Burak Özbagci, Yusuf Can, Fidel Tewolde



neufische Capstone Project - Gesellenstück



Project Overview

- Rossmann operates over 4,000 drug stores in 8 European countries
- Store managers are tasked with predicting their daily sales for up to six weeks in advance
- Store sales are influenced by many factors: promotion, holidays, seasonality...




This project aims to accurately predict sales for the upcoming six weeks



Data Description

Rossmann provided 3
raw data-sets

- 
- train.csv:
 - historical data including 'sales'
 - timeframe: January 1, 2013 - July 31, 2015
 - test.csv:
 - historical data excluding 'sales'
 - timeframe: August 1, 2015 - September 17, 2015
 - store.csv:
 - supplemental information about 1,115 stores



Evaluation Metric

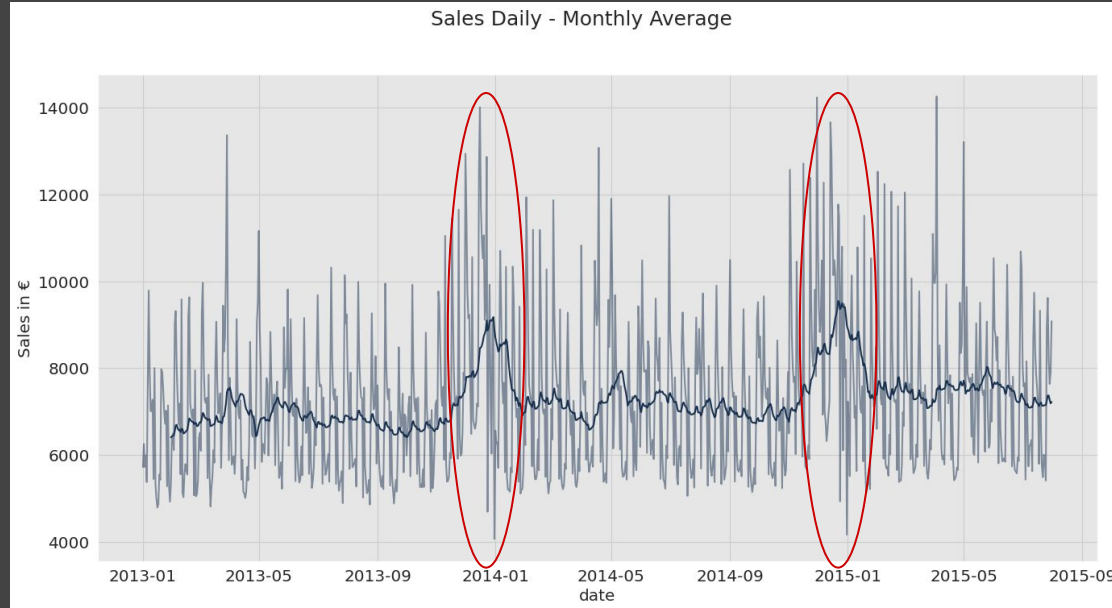
- Sales forecast for six-weeks
 - target timeframe in test.csv excludes 'sales', not possible to train on test-dataset
- Evaluation being done by minimizing the Root Mean Square Percentage Error - *RMSPE*



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

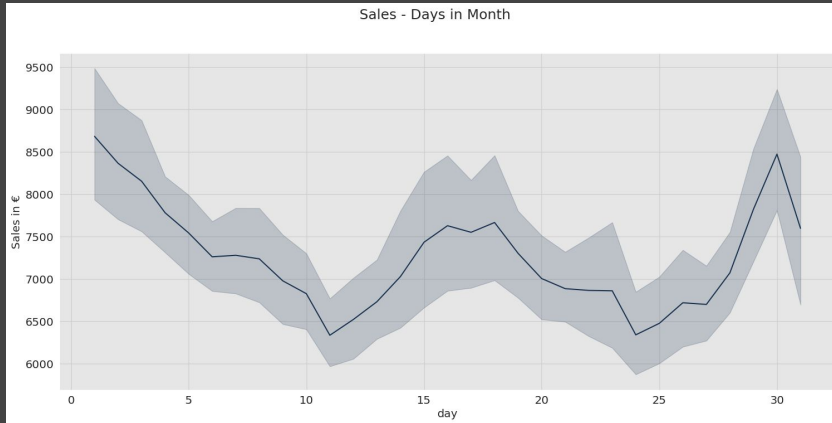
Visualisation of daily sales over three years

Sales increases by the end of each year (Christmas sales) followed by a steep decline before normalizing

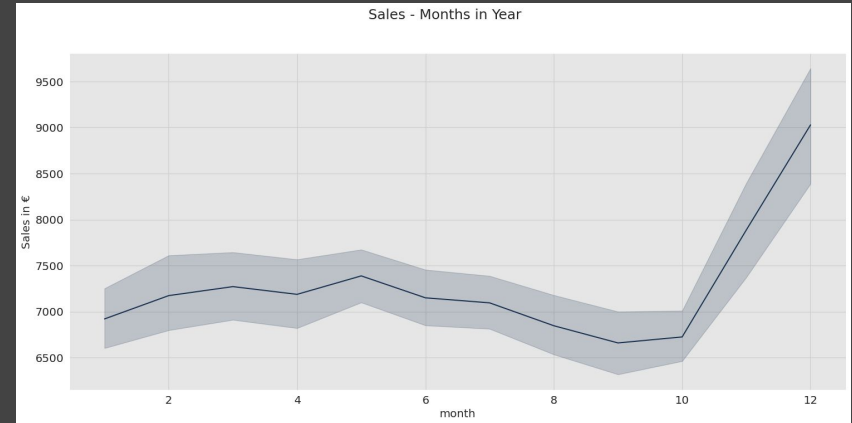


Visualisation of sales over a time period

Daily sales over a month:

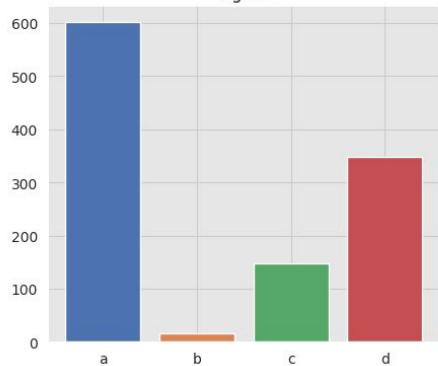


Monthly sales over a year:

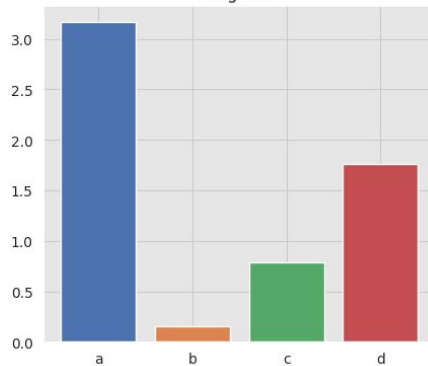


EDA Results

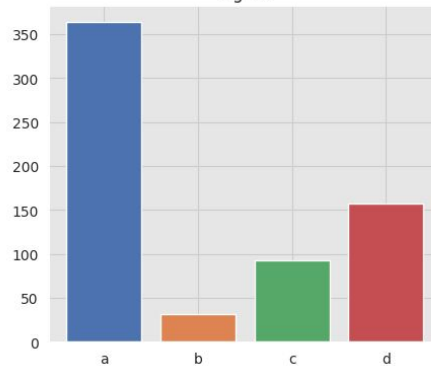
Number of Stores per Store Type
Fig 1.1



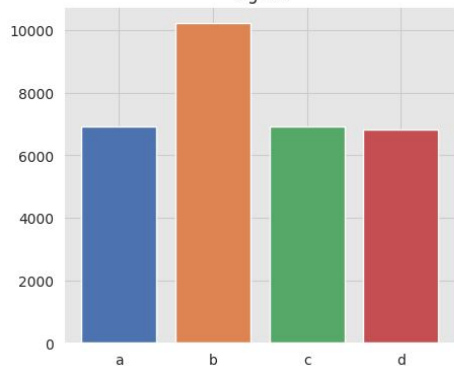
Total Sales per Store Type (in Billions)
Fig 1.2



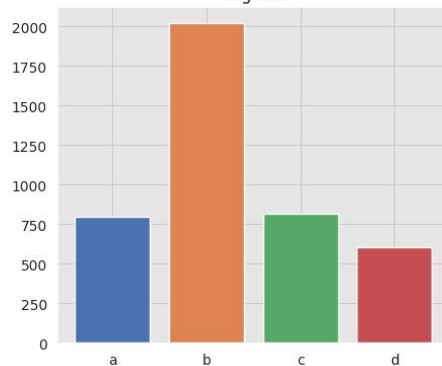
Total Number of Customers per Store Type (in Millions)
Fig 1.3



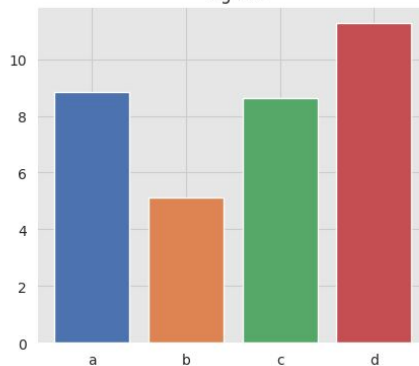
Average Sales per Store Type
Fig 1.4



Average Number of Customers per Store Type
Fig 1.5



Average Spending per Customer in each Store Type
Fig 1.6





Our Predictive Models

Ensemble Models

Random Forest Regressor

- Supervised Learning
- Decision Trees
- Trains models in isolation of one another

XG-Boost

- Extrem Gradient Boosting
- Trains models in succession
- Each iteration makes improvements

Neural Network Models

Dense Neural Network

- Fully connected layers
- Three hidden layers
- 512 nodes

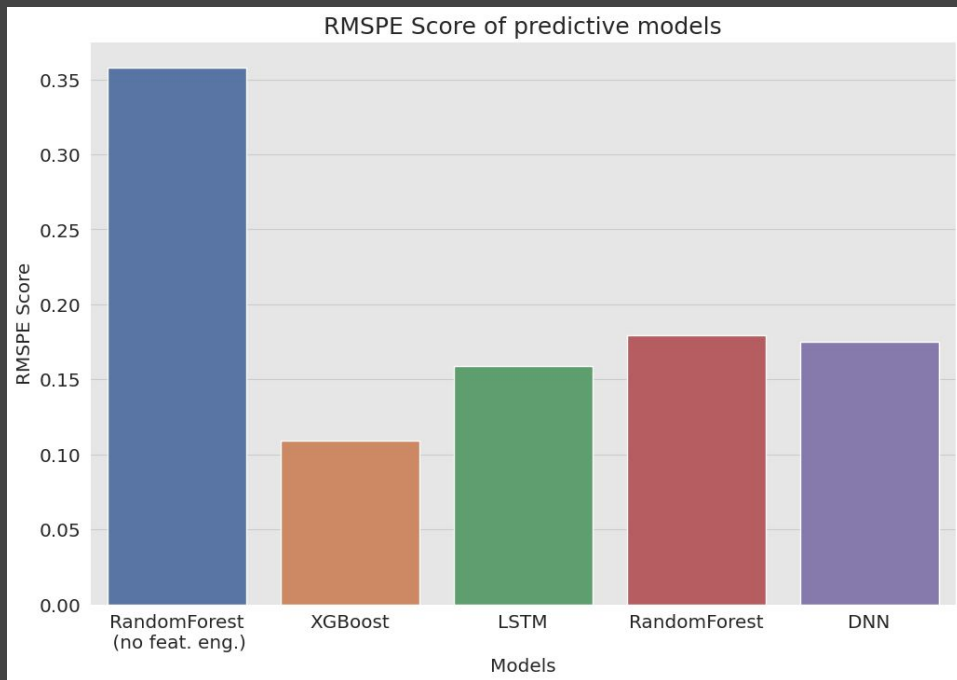
Neural Prophet

- Autoregressive (AR-) Model
- Time Series Model

LSTM

- Handles sequence dependent data
- Two LSTM layers
- 256 nodes

Predictive Model Results



Models	RMSPE Score	Train Duration
<i>RFR</i> (no feat. eng.)	0.357	< 3 min
<i>XGBoost</i>	0.109	~ 6 hours
<i>LSTM</i>	0.159	~ 8 hours (GPU)
<i>RFR</i>	0.179	< 3 min
<i>DNN</i>	0.175	~ 45 min (GPU)



Summary

- Out of over 3000 participants, we achieved 12th place with our best score
- RMSPE: 10.9% = by implication our predictions are 89,1% accurate
- Result could have been improved with external data (e.g. weather, geographical)
- Ensemble models outperform other methods
- Neural networks are robust estimators but have high cost



ROSSMANN Store Sales Challenge

Time Series Analysis for Sales Forecast

Burak Özbagci: <https://www.linkedin.com/in/burak-%C3%B6zbagci-75b532155>
Yusuf Can: <https://www.linkedin.com/in/yusuf-can-101282216>
Fidel Tewolde: <https://www.linkedin.com/in/fidel-tewolde>

