Rossmann Store Sales Prediction

Diandra MELO

SUMMARY

01 CONTEXT

Rossmann, Business Problem, Business Understanding

02 CHALLENGES

Main Problem, Possible Causes, Solutions

03 METHODOLOGY

Data Description, Hypothesis Creation, Exploratory Data Analysis, ML models

04 CONCLUSION

Business Performance, Machine Learning Model Performance

05 DEMONSTRATION

Model Deployment through Heroku, Telegram Bot

06 NEXT STEPS

Further actions to improve project

1

CONTEXT

ROSSMANN

drugstore chain with more than

3.000 stores

located in Europe

BUSINESS PROBLEM

Rossmann store managers are tasked with **predicting their daily sales** for up to **six weeks in advance**

BUSINESS UNDERSTANDING

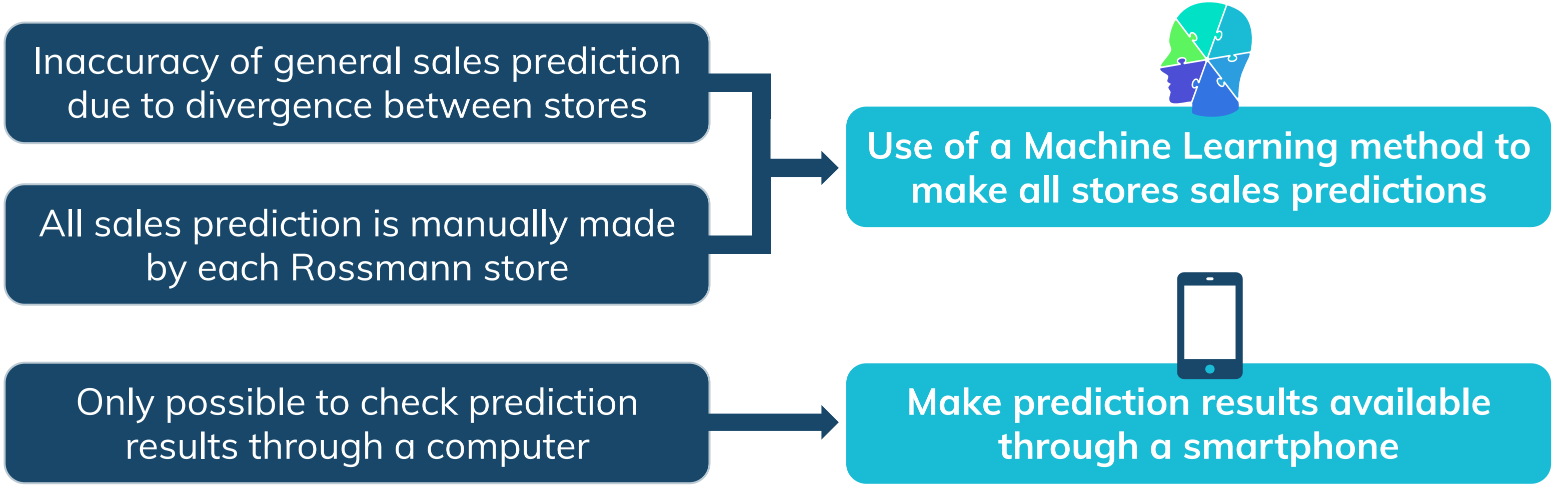
CFO wants to **invest individually in each store**, according to their incomes

2

CHALLENGES

budget definition for
individual investment of stores

PROBLEM



Inaccuracy of general sales prediction due to divergence between stores

All sales prediction is manually made by each Rossmann store

Only possible to check prediction results through a computer

Use of a Machine Learning method to make all stores sales predictions

Make prediction results available through a smartphone

3 METHODOLOGY

3.1 Data Description

Sales data for **1.115** Rossmann stores

Number of Columns: 18

Number of Rows: 1017209

DATE

Date of sales, holidays

STORE

Type, assortment

CUSTOMER

Quantity

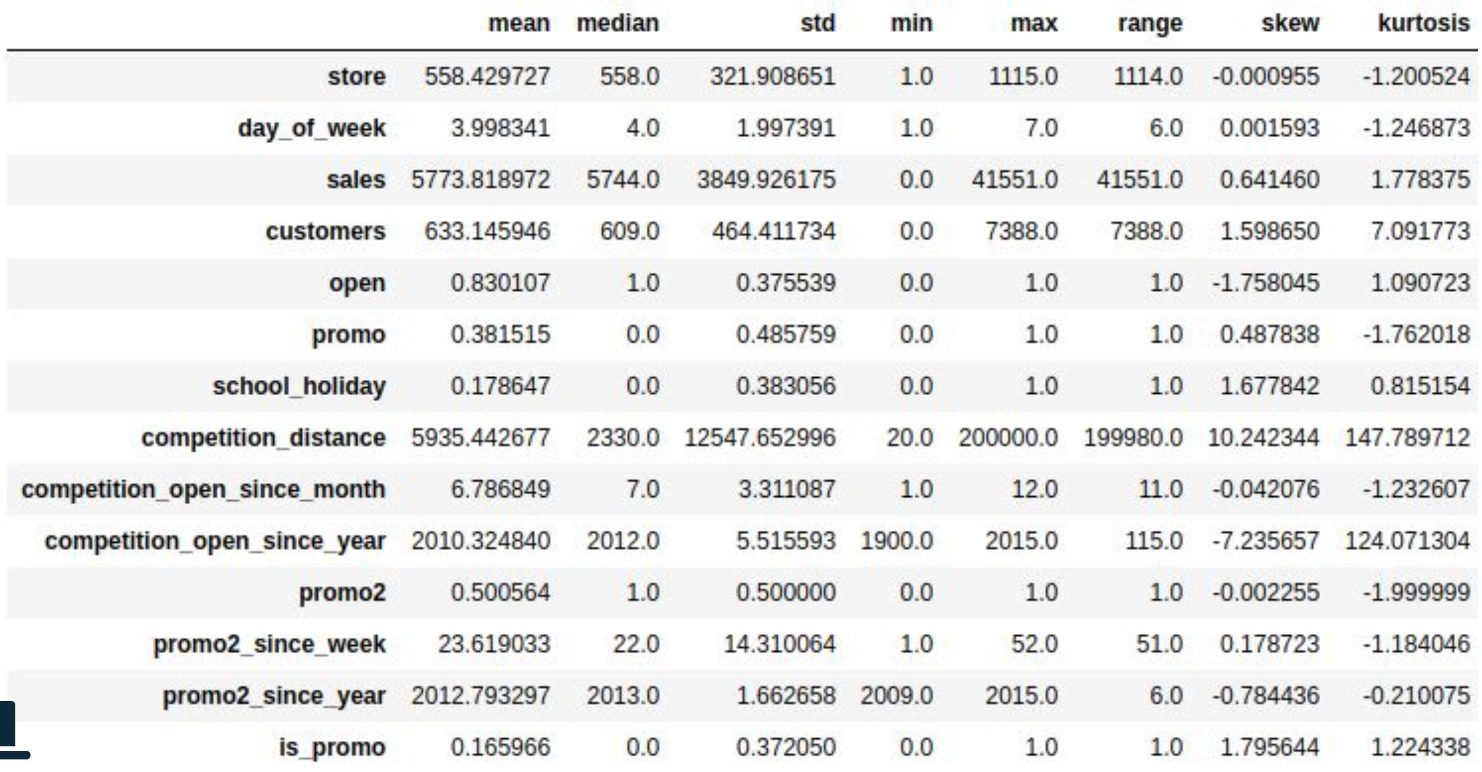
PROMOTIONS

Period, consecutive promo

COMPETITION

Distance, open since

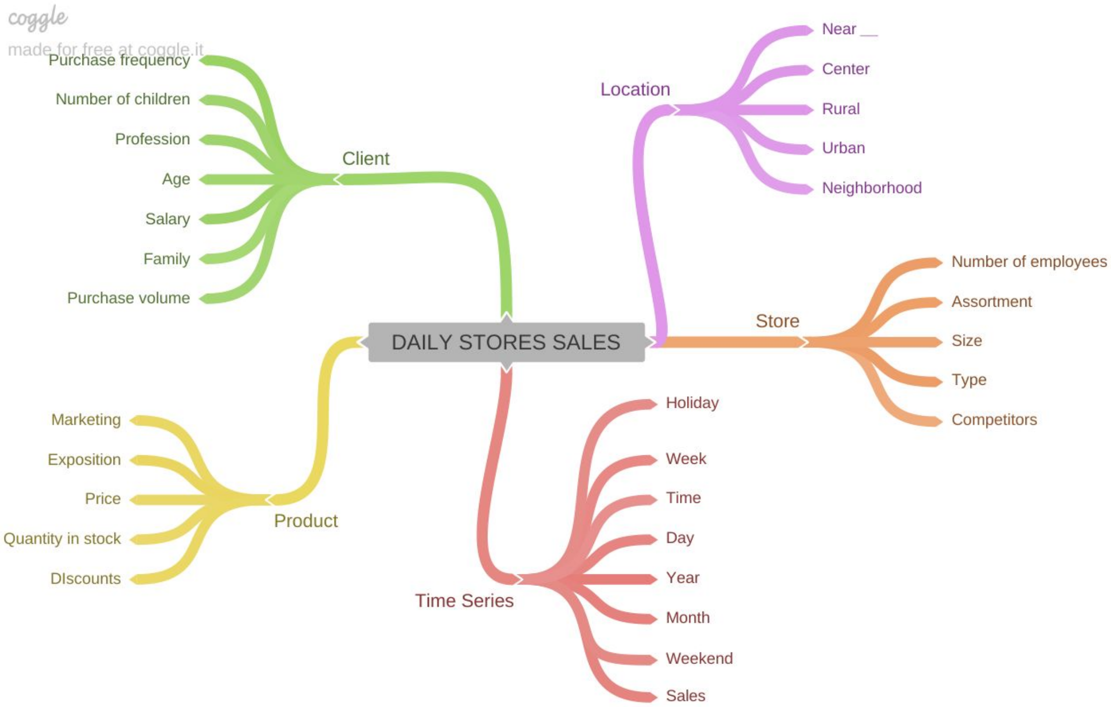
3.2 Descriptive Statistics



	mean	median	std	min	max	range	skew	kurtosis
store	558.429727	558.0	321.908651	1.0	1115.0	1114.0	-0.000955	-1.200524
day_of_week	3.998341	4.0	1.997391	1.0	7.0	6.0	0.001593	-1.246873
sales	5773.818972	5744.0	3849.926175	0.0	41551.0	41551.0	0.641460	1.778375
customers	633.145946	609.0	464.411734	0.0	7388.0	7388.0	1.598650	7.091773
open	0.830107	1.0	0.375539	0.0	1.0	1.0	-1.758045	1.090723
promo	0.381515	0.0	0.485759	0.0	1.0	1.0	0.487838	-1.762018
school_holiday	0.178647	0.0	0.383056	0.0	1.0	1.0	1.677842	0.815154
competition_distance	5935.442677	2330.0	12547.652996	20.0	200000.0	199980.0	10.242344	147.789712
competition_open_since_month	6.786849	7.0	3.311087	1.0	12.0	11.0	-0.042076	-1.232607
competition_open_since_year	2010.324840	2012.0	5.515593	1900.0	2015.0	115.0	-7.235657	124.071304
promo2	0.500564	1.0	0.500000	0.0	1.0	1.0	-0.002255	-1.999999
promo2_since_week	23.619033	22.0	14.310064	1.0	52.0	51.0	0.178723	-1.184046
promo2_since_year	2012.793297	2013.0	1.662658	2009.0	2015.0	6.0	-0.784436	-0.210075
is_promo	0.165966	0.0	0.372050	0.0	1.0	1.0	1.795644	1.224338

3.3 Hypothesis Creation

Hypothesis MindMap



coggle
made for free at coggle.it

DAILY STORES SALES

- Client
  - Purchase frequency
  - Number of children
  - Profession
  - Age
  - Salary
  - Family
  - Purchase volume
- Location
  - Near __
  - Center
  - Rural
  - Urban
  - Neighborhood
- Store
  - Number of employees
  - Assortment
  - Size
  - Type
  - Competitors
- Product
  - Marketing
  - Exposition
  - Price
  - Quantity in stock
  - DIscounts
- Time Series
  - Holiday
  - Week
  - Time
  - Day
  - Year
  - Month
  - Weekend
  - Sales

3.3 Hypothesis Creation

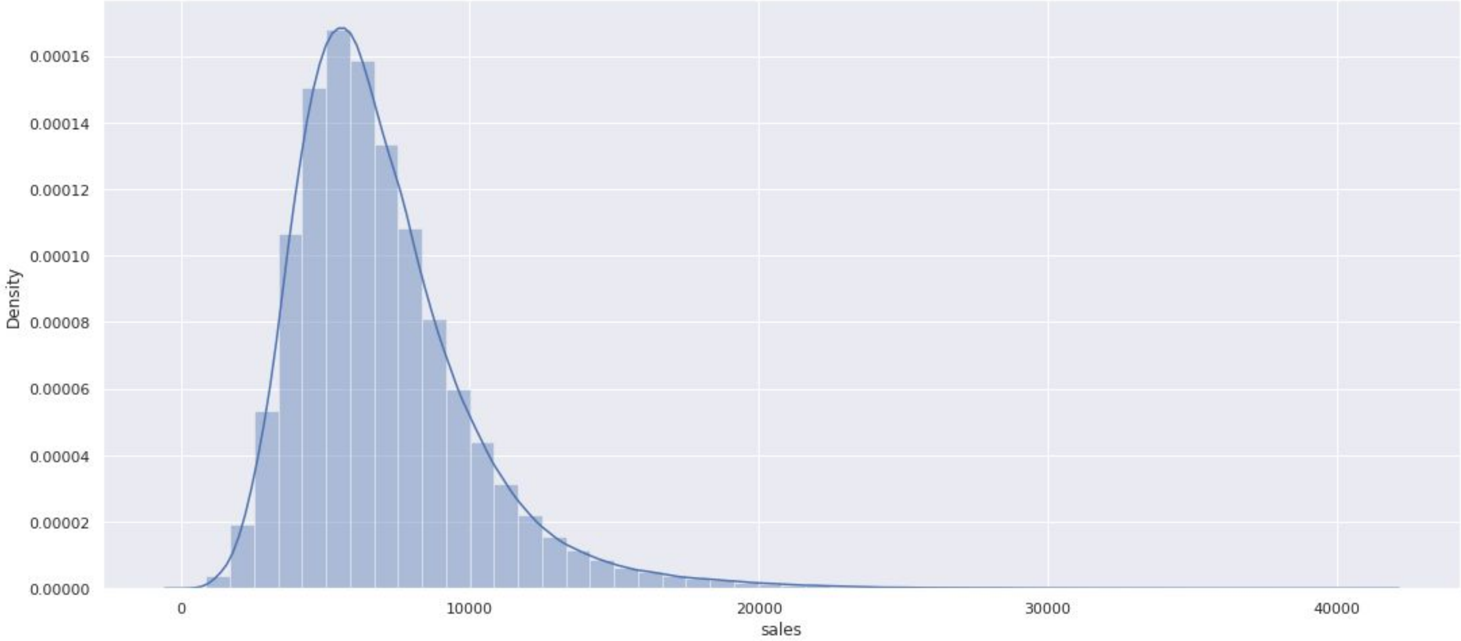
List of Hypothesis

1. Stores with **more diversified products** should sell **more**
2. Stores with **closer competitors** should sell **less**
3. Stores with **longer active promotions** should sell **more**
4. Stores with **more consecutive promotions** should sell **more**
5. Stores, during **school holidays**, should sell **more**
6. Stores, **in Christmas period**, should sell **more** than during other holidays
7. Stores, during **weekends**, should sell **less**
8. At the **first half of the month**, stores should sell **more**
9. In the **first semester of the year**, stores should sell **less**
10. **Throughout the years**, stores should sell **more**

3.4

Exploratory Data Analysis

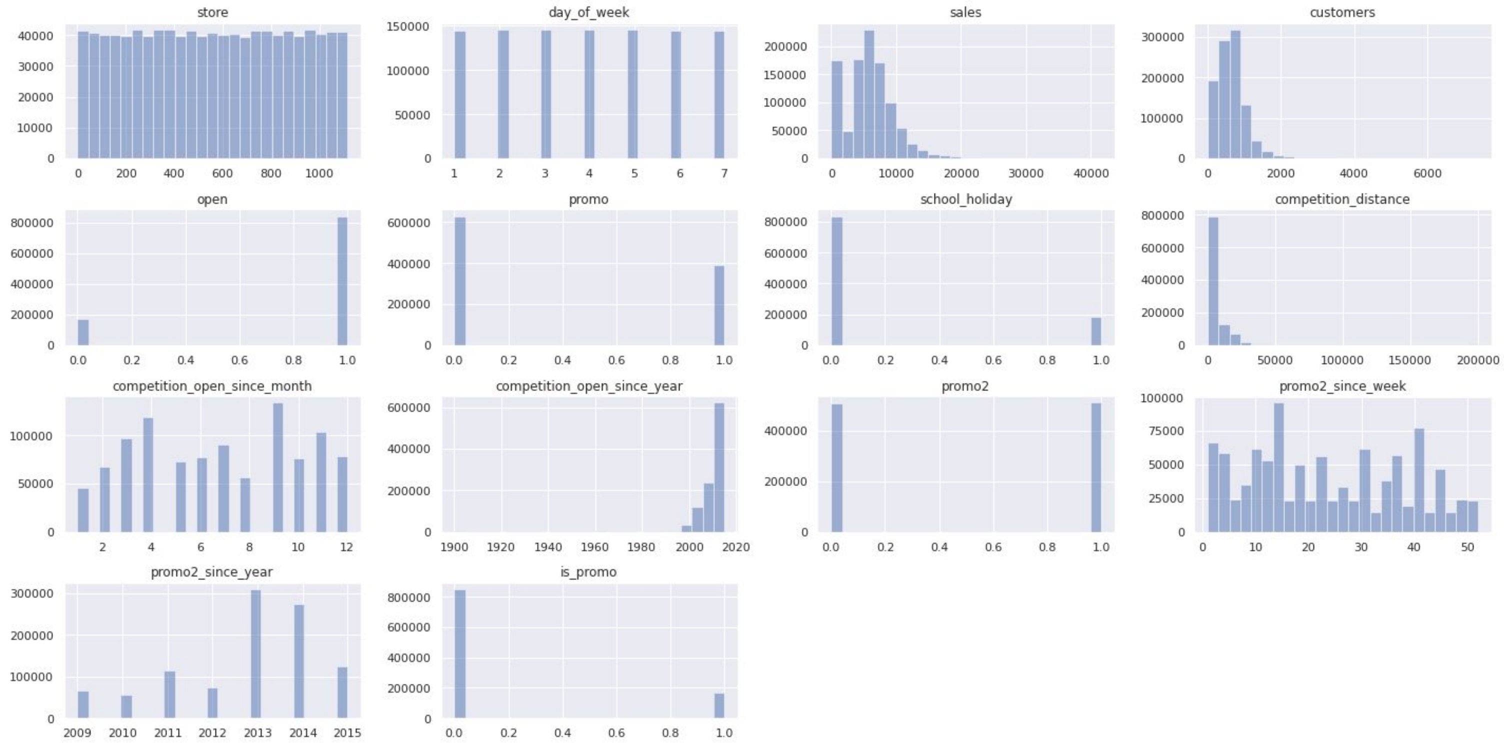
Univariate Analysis



3.4

Exploratory Data Analysis

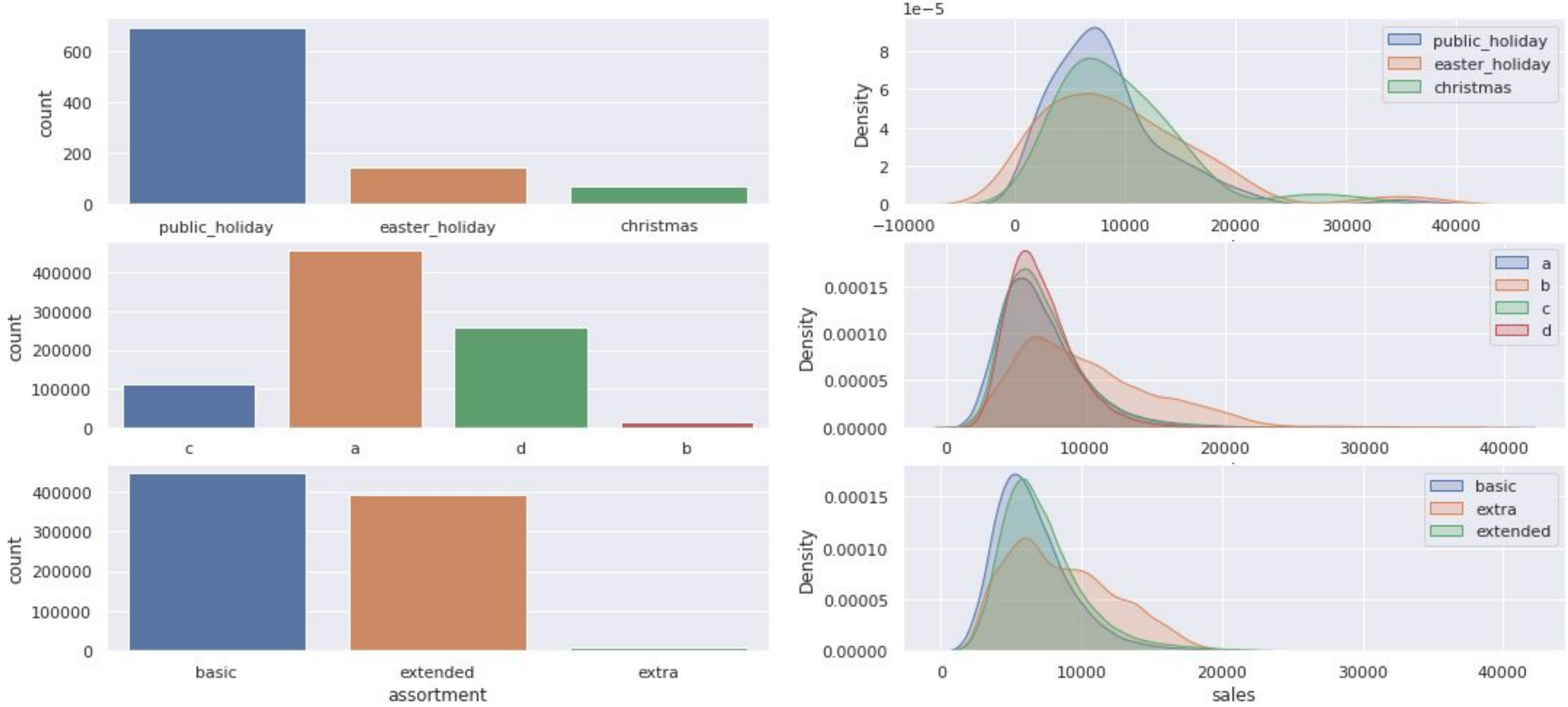
Univariate Analysis



3.4

Exploratory Data Analysis

Univariate Analysis

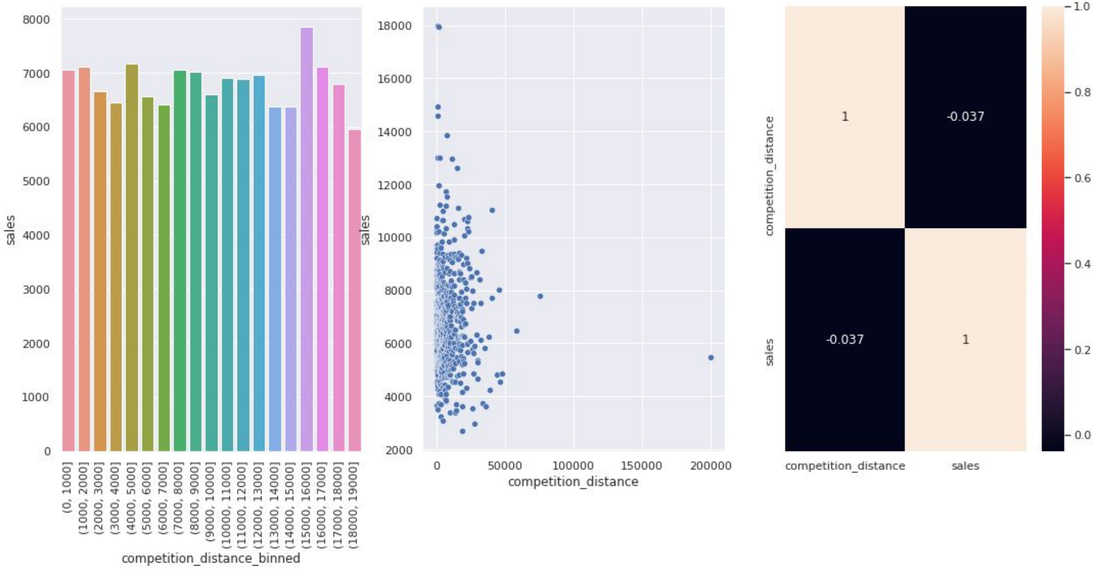


3.4

Exploratory Data Analysis

Hypothesis 02

Stores with closer competitors should sell less



This hypothesis turned to be **FALSE**.

Mean of sales does not vary much: in fact, the highest mean value is found for higher distances.

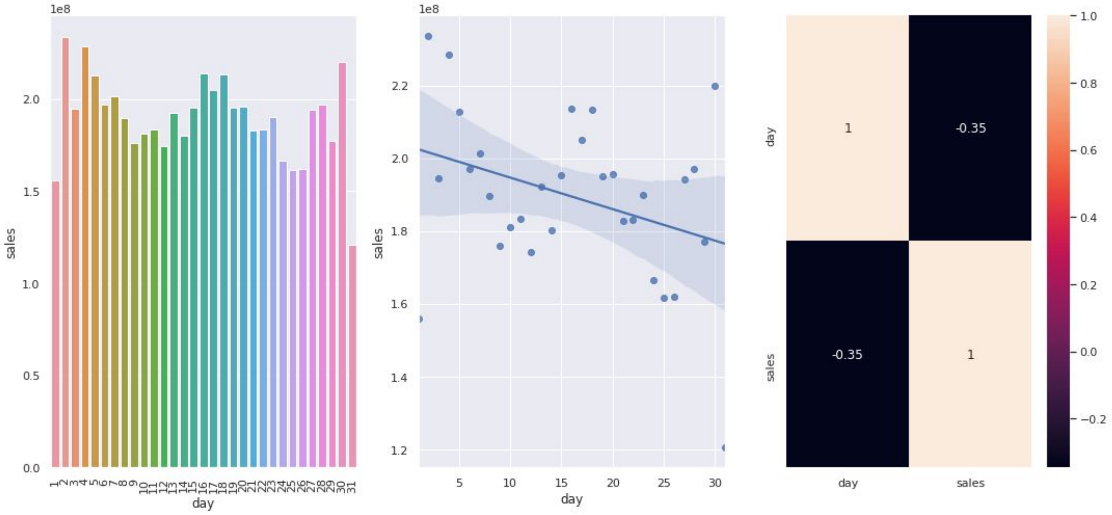
Not a strong correlation (heat map).

3.4

Exploratory Data Analysis

Hypothesis 08

At the first half of the month, stores should sell more



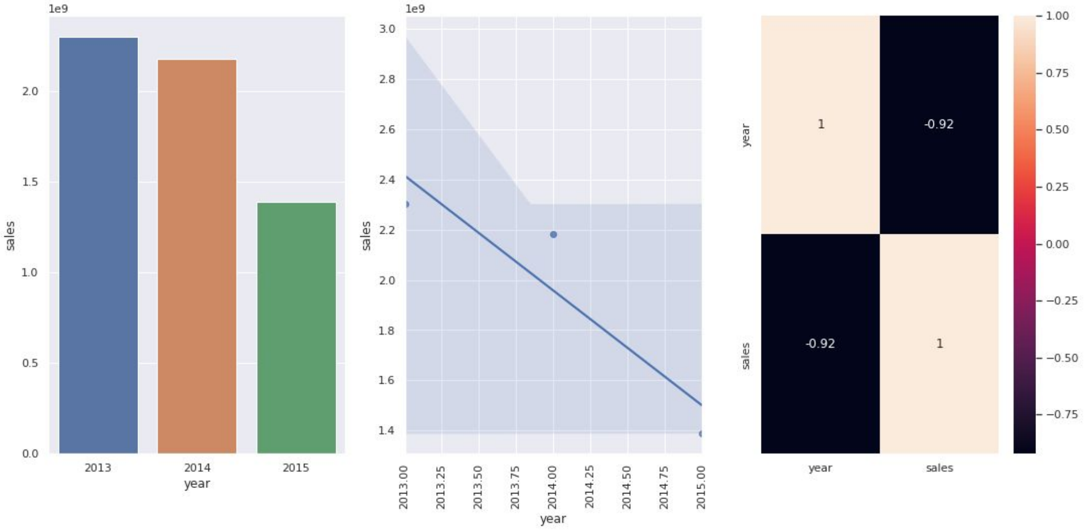
This hypothesis turned to be **TRUE**.

◇ The linear regression slope exhibits well this tendency.

3.4 Exploratory Data Analysis

Hypothesis 10

Throughout the years, stores should sell more



This hypothesis turned to be **FALSE**.

◇ Even though data of 2015 is incomplete (until July), it was considered to observe the tendency of sales reduction throughout years.

3.4 Exploratory Data Analysis

Hypothesis Final Table

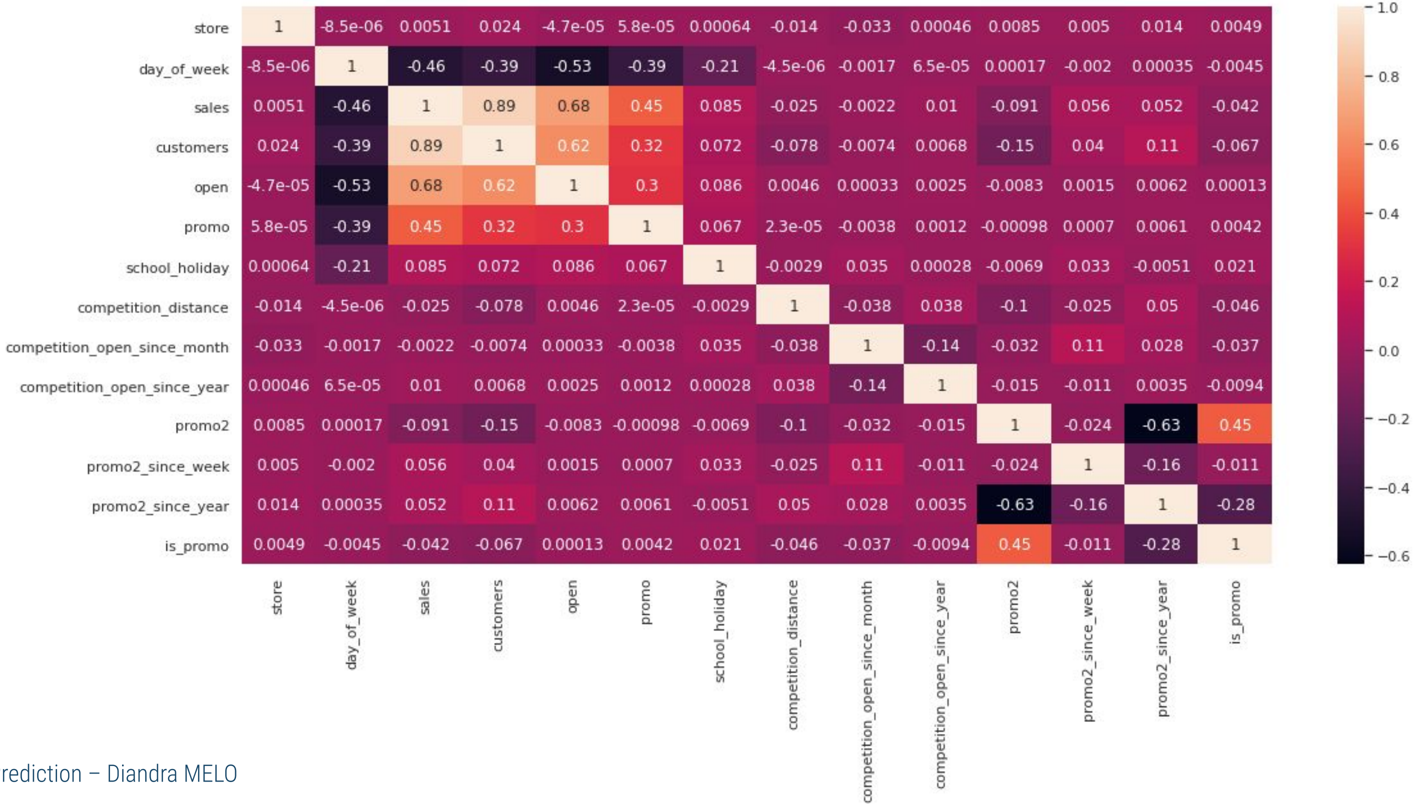
Hypothesis	Conclusion	Relevance
H1 – Stores with more diversified products should sell more	TRUE	Medium
H2 – Stores with closer competitors should sell less	FALSE	Low
H3 – Stores with longer active promotions should sell more	TRUE	Medium
H4 – Stores with more consecutive promotions should sell more	FALSE	Low
H5 – Stores, during school holidays, should sell more	TRUE	Low
H6 – Stores, in Christmas period, should sell more than during other holidays	FALSE	Medium
H7 – Stores, during weekends, should sell less	TRUE	High
H8 – At the first half of the month, stores should sell more	TRUE	High
H9 – In the first semester of the year, stores should sell less	TRUE	High
H10 – Throughout the years, stores should sell more	FALSE	High

3.4

Exploratory Data Analysis

Multivariate Analysis

Numerical Attributes



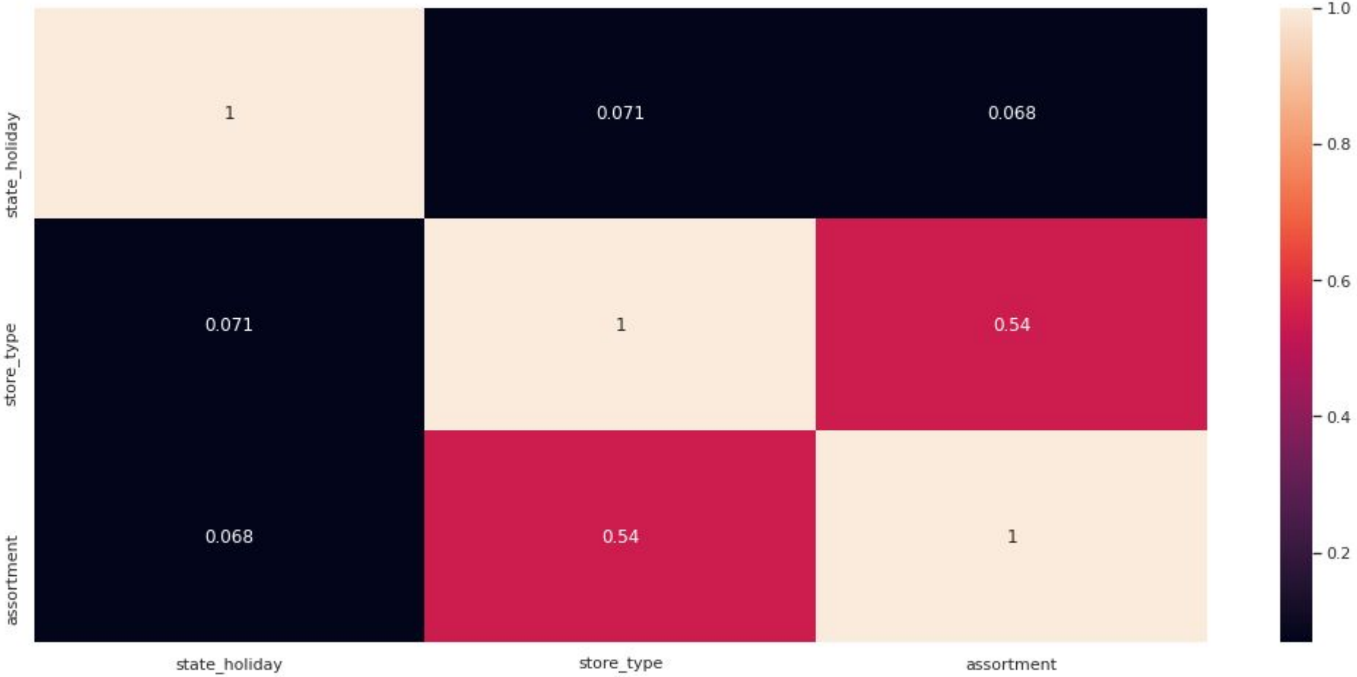
	store	day_of_week	sales	customers	open	promo	school_holiday	competition_distance	competition_open_since_month	competition_open_since_year	promo2	promo2_since_week	promo2_since_year	is_promo
store	1	-8.5e-06	0.0051	0.024	-4.7e-05	5.8e-05	0.00064	-0.014	-0.033	0.00046	0.0085	0.005	0.014	0.0049
day_of_week	-8.5e-06	1	-0.46	-0.39	-0.53	-0.39	-0.21	-4.5e-06	-0.0017	6.5e-05	0.00017	-0.002	0.00035	-0.0045
sales	0.0051	-0.46	1	0.89	0.68	0.45	0.085	-0.025	-0.0022	0.01	-0.091	0.056	0.052	-0.042
customers	0.024	-0.39	0.89	1	0.62	0.32	0.072	-0.078	-0.0074	0.0068	-0.15	0.04	0.11	-0.067
open	-4.7e-05	-0.53	0.68	0.62	1	0.3	0.086	0.0046	0.00033	0.0025	-0.0083	0.0015	0.0062	0.00013
promo	5.8e-05	-0.39	0.45	0.32	0.3	1	0.067	2.3e-05	-0.0038	0.0012	-0.00098	0.0007	0.0061	0.0042
school_holiday	0.00064	-0.21	0.085	0.072	0.086	0.067	1	-0.0029	0.035	0.00028	-0.0069	0.033	-0.0051	0.021
competition_distance	-0.014	-4.5e-06	-0.025	-0.078	0.0046	2.3e-05	-0.0029	1	-0.038	0.038	-0.1	-0.025	0.05	-0.046
competition_open_since_month	-0.033	-0.0017	-0.0022	-0.0074	0.00033	-0.0038	0.035	-0.038	1	-0.14	-0.032	0.11	0.028	-0.037
competition_open_since_year	0.00046	6.5e-05	0.01	0.0068	0.0025	0.0012	0.00028	0.038	-0.14	1	-0.015	-0.011	0.0035	-0.0094
promo2	0.0085	0.00017	-0.091	-0.15	-0.0083	-0.00098	-0.0069	-0.1	-0.032	-0.015	1	-0.024	-0.63	0.45
promo2_since_week	0.005	-0.002	0.056	0.04	0.0015	0.0007	0.033	-0.025	0.11	-0.011	-0.024	1	-0.16	-0.011
promo2_since_year	0.014	0.00035	0.052	0.11	0.0062	0.0061	-0.0051	0.05	0.028	0.0035	-0.63	-0.16	1	-0.28
is_promo	0.0049	-0.0045	-0.042	-0.067	0.00013	0.0042	0.021	-0.046	-0.037	-0.0094	0.45	-0.011	-0.28	1

3.4

Exploratory Data Analysis

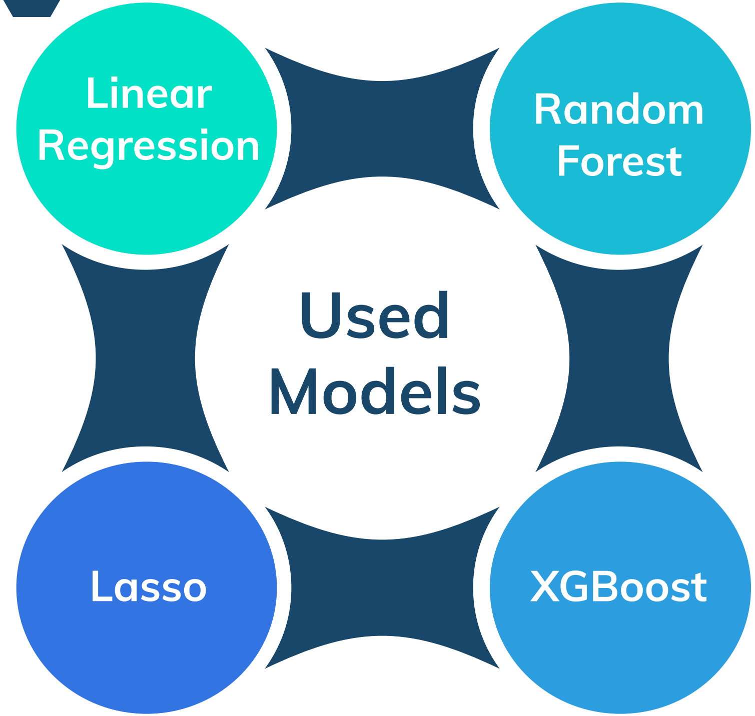
Multivariate Analysis

Categorical Attributes



	state_holiday	store_type	assortment
state_holiday	1	0.071	0.068
store_type	0.071	1	0.54
assortment	0.068	0.54	1

3.5 Machine Learning Modelling



Linear Regression

Random Forest

Used Models

Lasso

XGBoost

Model's Performance

Model Name	MAE CV	MAPE CV	RMSE CV
Linear Regression	2081.73 +/- 295.63	0.3 +/- 0.02	2952.52 +/- 468.37
Lasso	2116.38 +/- 341.5	0.29 +/- 0.01	3057.75 +/- 504.26
RF Regressor	837.68 +/- 219.1	0.12 +/- 0.02	1256.08 +/- 320.36
XGBoost Regressor	1030.28 +/- 167.19	0.14 +/- 0.02	1478.26 +/- 229.79

MODEL'S CHOICE

XGBoost Regressor

◇ Lighter than RF Regressor, with not much difference in error

4

CONCLUSION

4.1 Business Performance

285,860,497.77 $

Total Stores Sales Prediction for the next 6 weeks

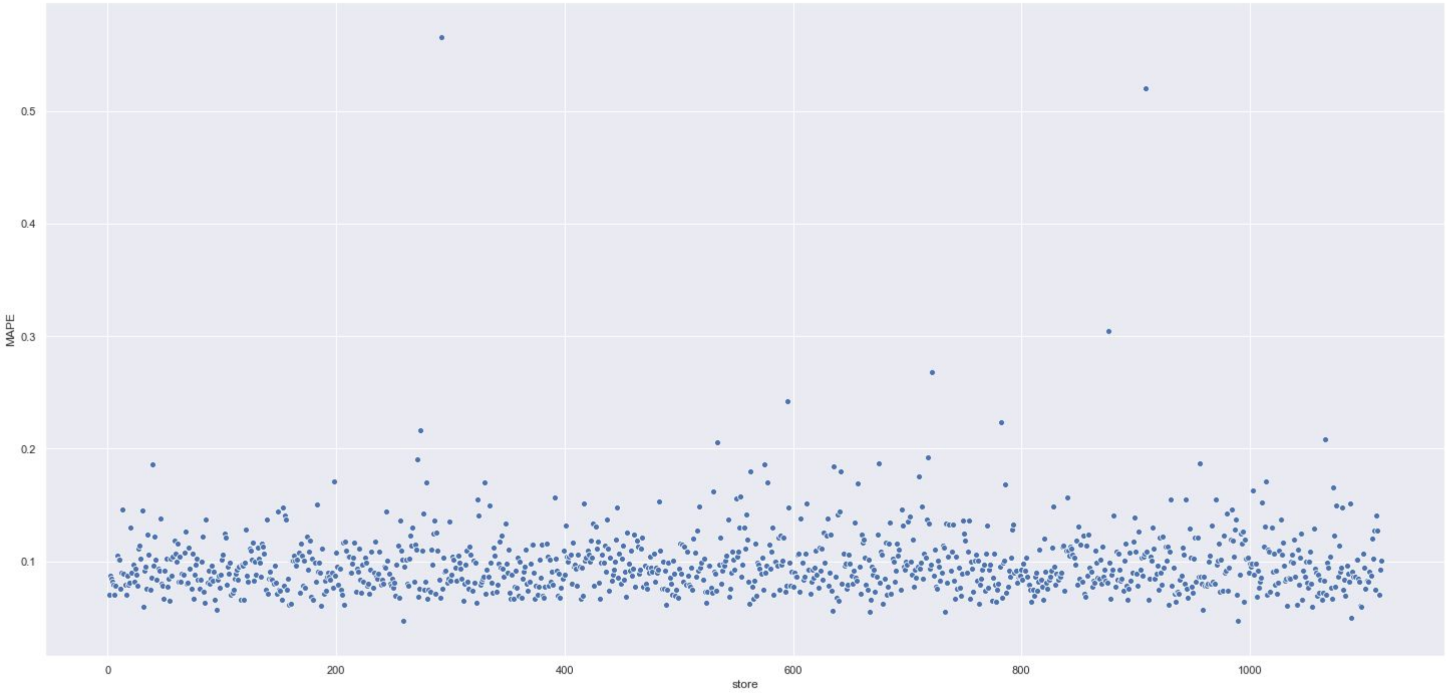
± 745,482.07 $

Margin of error (Worst/Best Scenario)

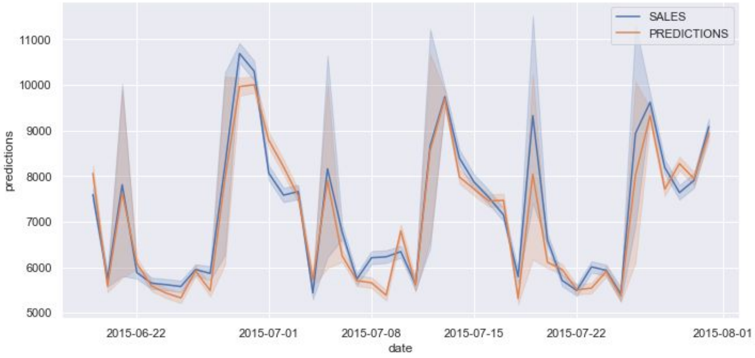
< 10 %

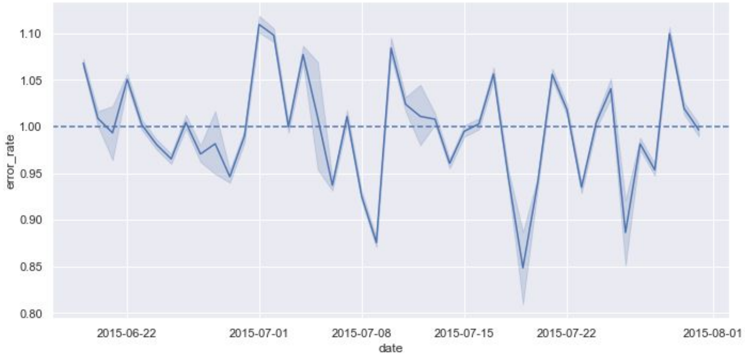
Average percentage error of the stores sales

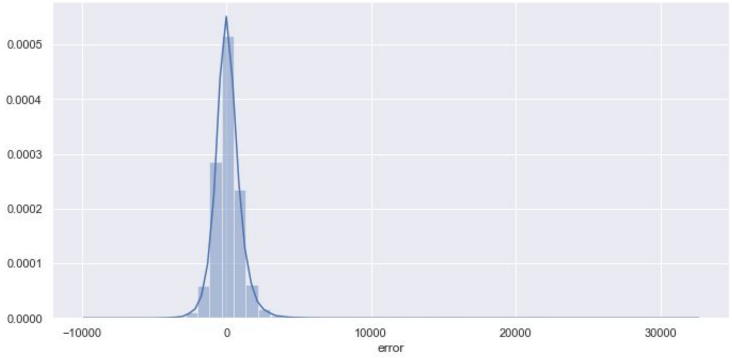
4.1 Business Performance

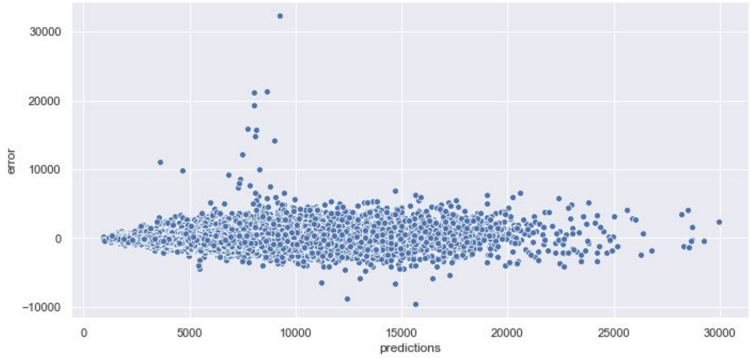


4.2 Machine Learning Performance







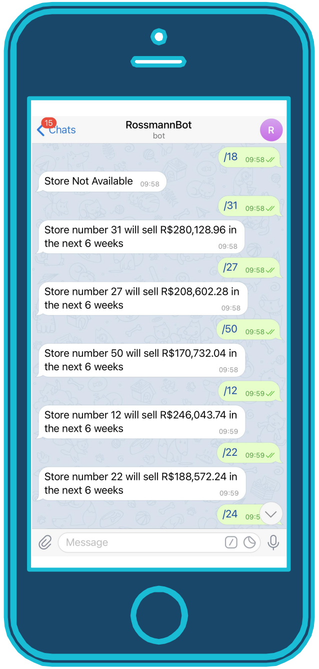


5

DEMONSTRATION

Model Deployment - Heroku

- Telegram Bot



/18 09:58

Store Not Available 09:58

/31 09:58

Store number 31 will sell R$280,128.96 in the next 6 weeks 09:58

/27 09:58

Store number 27 will sell R$208,602.28 in the next 6 weeks 09:58

/50 09:58

Store number 50 will sell R$170,732.04 in the next 6 weeks 09:58

/12 09:59

Store number 12 will sell R$246,043.74 in the next 6 weeks 09:59

/22 09:59

Store number 22 will sell R$188,572.24 in the next 6 weeks 09:59

/24 09:5

6. Next Steps

- Improve Model’s performance;
- Test usability of deployment;
- Check for other ML models that could fit well into this study case.

THANKS!

Any questions?

You can find me at:

- LinkedIn: Diandra Melo
- GitHub: @diandramelo
- Gmail: dcxsmelo@gmail.com

Credits

I would like to give an special thanks to **Meigarom Lopes** for providing the orientation needed for me to achieve these results, as well as for improving my Data Science knowledge throughout the course ***Data Science in Production***.

- Forecast sales competition *Rossmann Store Sales* from kaggle.com
- Presentation template by SlidesCarnival