PIZZA RESTAURANT DATA ANALYSIS PROJECT

Ironhack week 7: Machine learning

Contributors:

Iván
Jorge
Rajashree
Matt

Perfectly·Distributed·Pizza

Project workflow

Data cleaning + preparation

Time series prediction

Categorical

Insight generation + plotting

Conclusions

Business case

- Predict peak times to optimise staffing
  - Track seasonality in different time periods: daily, weekly, monthly and yearly
  - Adjust staffing levels to maximise output at peak times and utilisation during troughs
  - Use seasonal insights to generate ideas for promotional activity
- Predict customer choices to gain insights for ingredient purchasing
  - Predict purchasing patterns in type, name, size and price to enable smart ordering practices for ingredients.
  - Give insights about the accuracy of the model for the different variables

Perfectly·Distributed·Pizza

SECTION

Optimising staffing

Using Meta’s Prophet model to forecast time series data

Perfectly·Distributed·Pizza

Models Used

- Prophet (Meta)

Meta

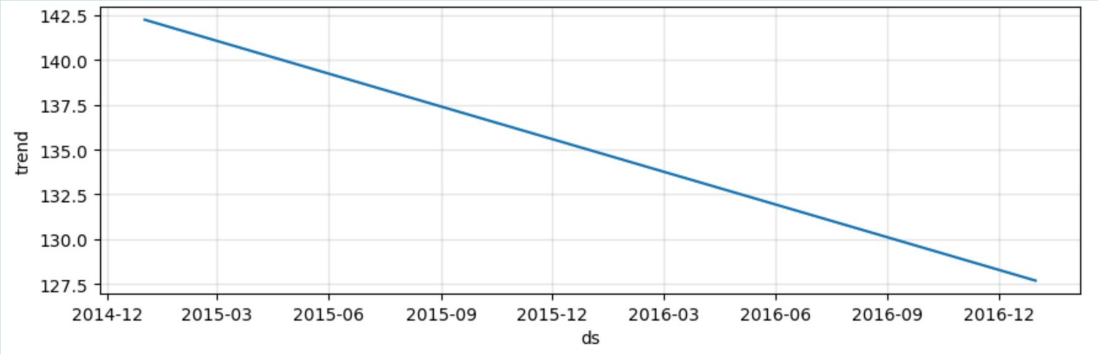
Functions

- Sklearn style .fit/.predict
- Seasonality forecasts
- Cross validation for MAE values

Perfectly·Distributed·Pizza

Forecasting daily orders

- Concerning downward trend in order volume across the year



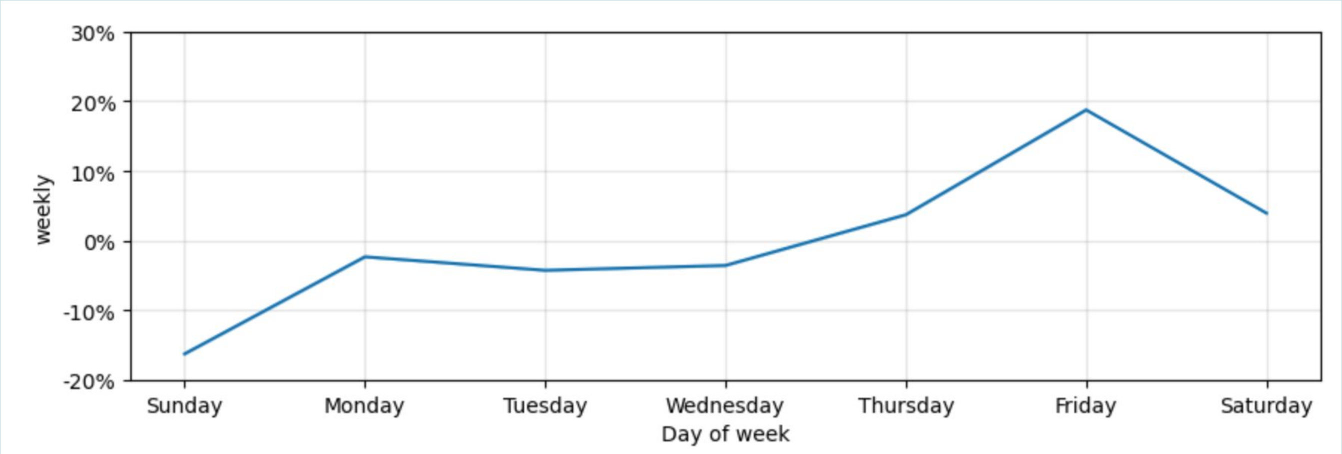
- Able to forecast future order volumes with some accuracy:
  - 30 days: MAE 18.6 (13.4%)
  - 60 days: MAE 25.0 (18.1%)
  - 90 days: MAE 33.8 (24.4%)

Mean daily orders: 138

Perfectly·Distributed·Pizza

Tracking weekly patterns

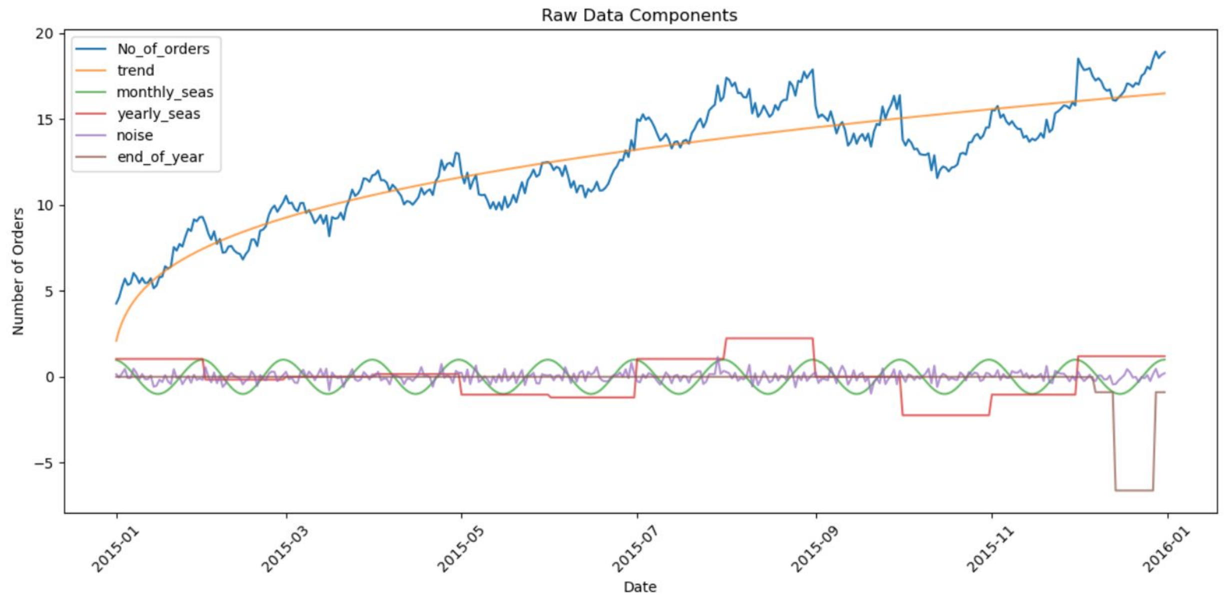
- Natural increase of 15% in order volume for “pre-weekend” period (Thursday-Saturday)



Perfectly·Distributed·Pizza

Monthly seasonality

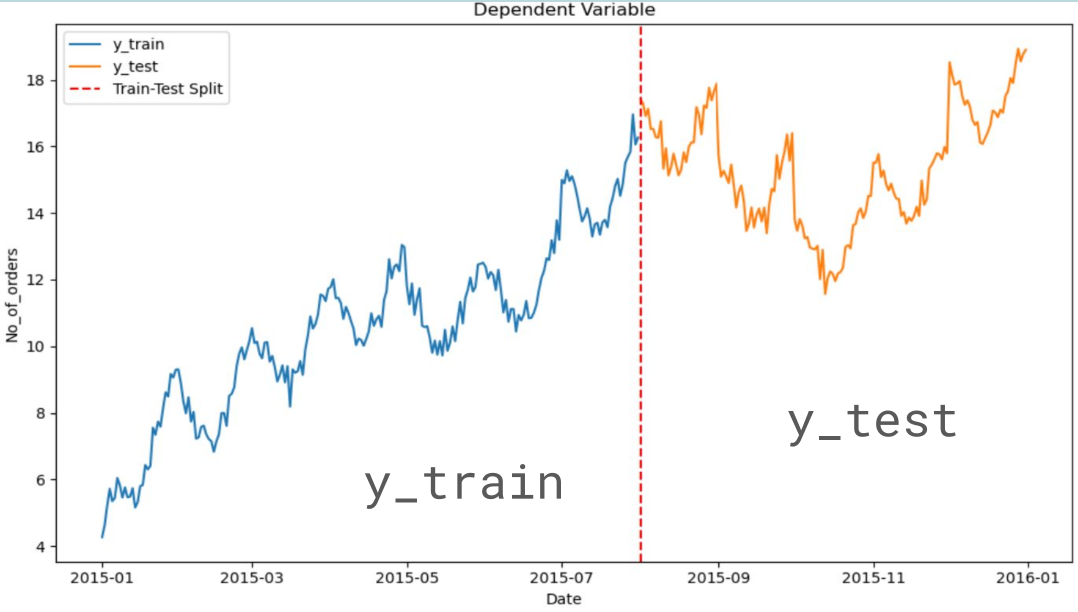
- Distinct seasonal peaks in sales—particularly in August and towards the end of the year



Perfectly·Distributed·Pizza

Monthly seasonality

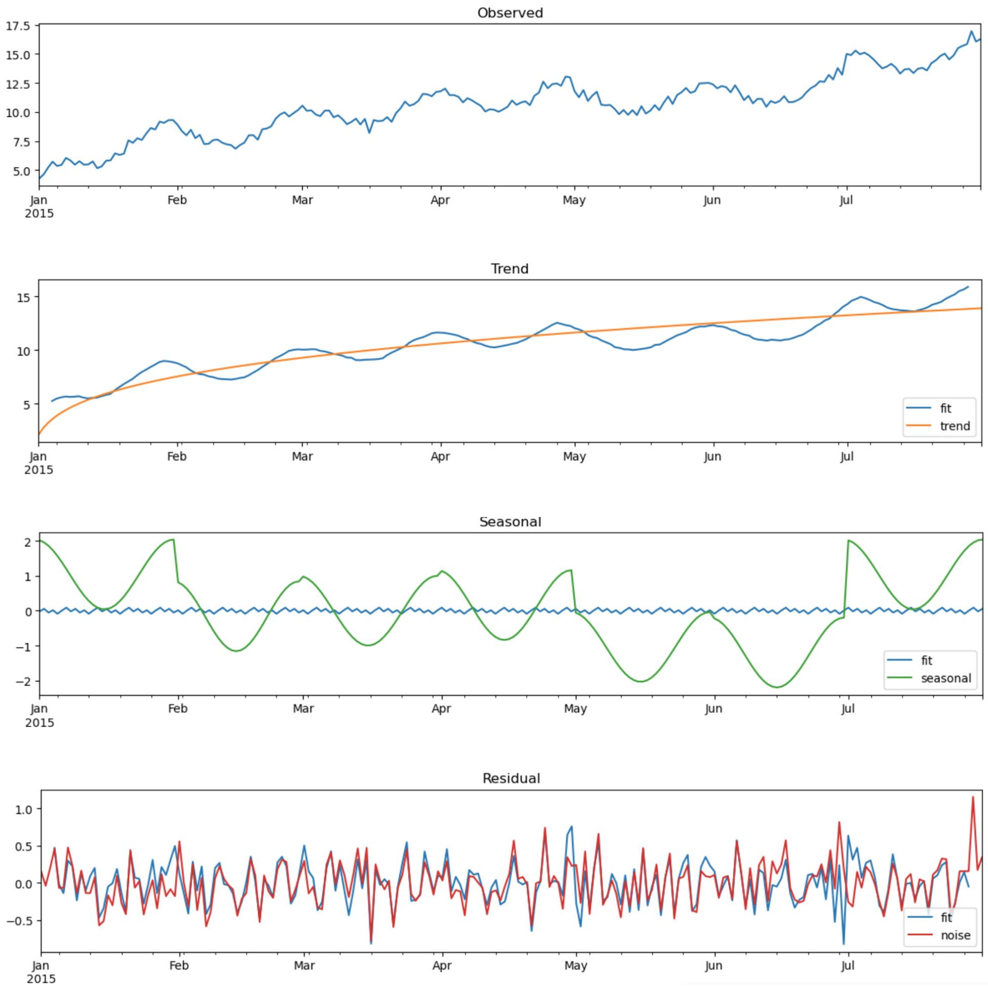
- Evaluating our model : threshold “2015-08-01”



Perfectly·Distributed·Pizza

Time Series Decomposition

Because time series often exhibit different patterns, **decomposition** can be used which breaks the time series down into **four parts**: observed data, trend, seasonal and remainder/residual.



SECTION

Mapping customer choices

- Categorical data transformation
- **Comparing methods**: KNN, Bagging, RFM, AdaBoost, Gradient Boost
- **Goal**: get the most accurate method/model
- Most optimal **hyperparameter**

Perfectly·Distributed·Pizza

Categorical Data Transformation

Methodologies used:

- One Hot Encoder to dummify categorical columns’ values 1 by 1
- Dictionaries and the use of map in pizza columns to values

[size]	[type]	[name]
S	Veggie	Margerita
M	Classic	Calabrese
L	Supreme	Pepperoni
XL	Chicken	[...]
XXL		

One Hot Encoding / Dictionary mapping

Perfectly·Distributed·Pizza

Findings

Using OneHotEncoder and then applying the k-nearest neighbors (KNN) algorithm to predict **pizza sales patterns**.

- Column [name] had an R^2 score of **0.23**.
    - **Insight**: [name] of pizza had low predictive power.
- Column [size] had an R^2 score of 1.00.
    - **Insight:** The [size] variable has moderate predictive power in this dataset which is close to 1. When we calculated size without relation to **pizza price** R^2 was closer to **0.47** so its very dependent on this.
- Column [type] had a perfect score of **1.00**
    - **Insight:** This is unusual and could be a sign of overfitting.

Perfectly·Distributed·Pizza

Models Used

- K-Nearest Neighbors
- Linear Regression
- Decision Trees

Ensemble Techniques

- Bagging and Pasting
- Random Forest Method
- Gradient Boosting
- Ada boost

Perfectly·Distributed·Pizza

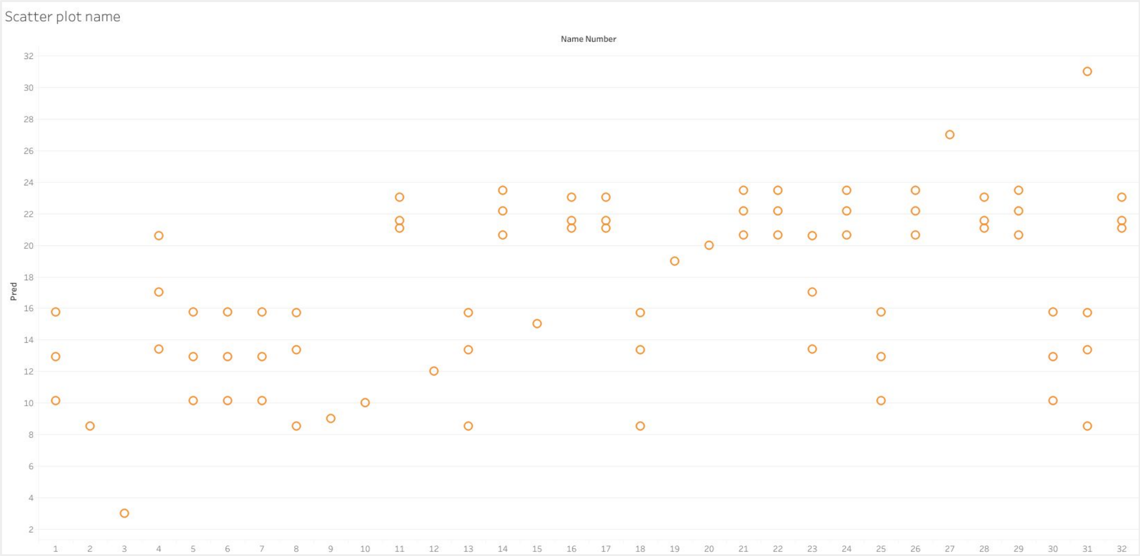
After comparing we found...

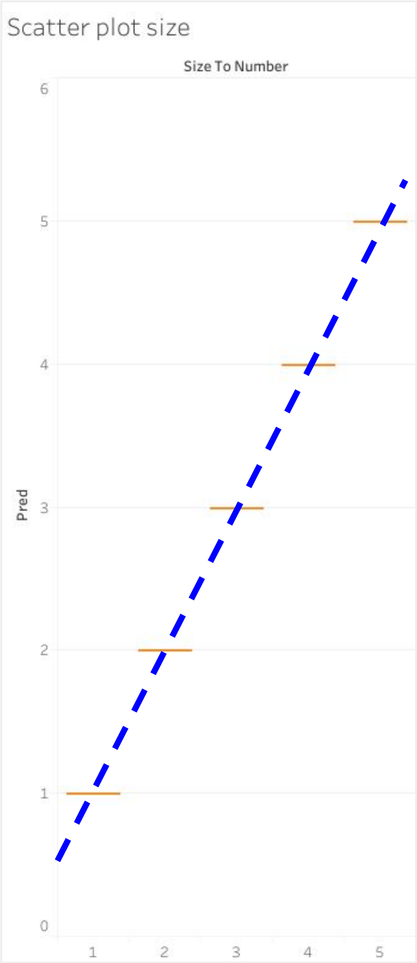
- **Random forest** was the best **ensemble method** according to the R2 score but it was still near to 0.36 for column **name** but near to 1.00 for [size] and [type].

- After several tries with the **Grid Search model** we have estimated that the optimal “**depth**" is 10 and "**max leaf nodes**" is 40.

Perfectly·Distributed·Pizza

Plotting [size] and [name]





Correlation

Biggest correlation in price with [size] and [type]

Almost no relation between the variables so probably will have a low R^2 value

Heatmap Correlation

Variables	Name Number	Price	Size To Number	Type To Number	Weekday Nº
name_number	1,0000	0,0853	0,1074	0,1471	0,0002
price	0,0853	1,0000	0,8843	0,2906	[illegible]
size_to_number	0,1074	0,8843	1,0000	0,0758	0,0079
type_to_number	0,1471	0,2906	0,0758	1,0000	[illegible]
weekday nº	0,0002	0,0100	0,0079	0,0032	1,0000

SECTION

Conclusions

How can Perfectly Distributed Pizza use these insights to generate more profit in 2016?

Perfectly·Distributed·Pizza

Optimise staffing

Time series predictions suggest that PDP's order volume will continue to decrease through 2016

Weekly seasonality makes a clear case for additional staffing on Thursday, Friday and Saturdays

Underperformance in June and October suggest indicative windows for promotional activity

Perfectly·Distributed·Pizza

Smart inventory management

Pizza dough predictions would make PDP able to optimize its inventory

However, modelling demand by [name] showed unpredictable demand as demonstrated by the low R^2 value

Categorical columns also showed limited correlation suggesting each should be analysed individually by volume

Perfectly·Distributed·Pizza

Thank you

Have a slice on us!

Perfectly·Distributed·Pizza