UNIVERSITÀ DEGLI STUDI DI CAGLIARI
FACOLTÀ DI SCIENZE ECONOMICHE GIURIDICHE E POLITICHE

CORSO DI LAUREA IN DATA SCIENCE, BUSINESS ANALYTICS E INNOVAZIONE

Analisi e previsione di serie storiche: confronto tra i modelli di statistical learning ARIMA, LSTM, Random forest e Boosting, applicati al caso delle vendite della Corporación Favorita.

Data Science, Business Analytics e Innovazione
Laurea Magistrale

Tesi di Laurea di: Andrea Corongiu

Relatore: Prof. Marco Ortu

Argomento della tesi

01

Analisi e **previsione** di **sei serie storiche** delle **vendite** giornaliere di **tre categorie di prodotto**, in **due punti vendita** della Corporación Favorita.

Rilevanza:

- **Task** frequente per un **data scientist**;
- **Ambito** di applicazione: **Ottimizzazione** della **gestione** delle **risorse** per aumentare **redditività**.

previsioni attendibili sono fondamentali durante la fase di **programmazione degli approvvigionamenti**;

CORPORACIÓN FAVORITA

Il dataset

05

Il dataset estrapolato da Kaggle è composto da **sei serie storiche**, quelle **più** corpose in termini di **unità vendute**.

Gli **store 44 e 45** si trovano a Quito, la capitale dell'Ecuador.

Le **categorie** di prodotto sono tre:

01 | **BEVERAGES**

02 | **GROCERY I**

03 | **PRODUCE**

	44 BEVERAGES	44 GROCERY I	44 PRODUCE	45 BEVERAGES	45 GROCERY I	45 PRODUCE
2013-01-01	5466.00	10686.00	57.00	4070.00	11422.00	47.00
2013-01-02	5466.00	10686.00	57.00	4070.00	11422.00	47.00
2013-01-03	3718.00	7342.00	57.00	2526.00	6841.00	47.00
2013-01-04	4112.00	7250.00	57.00	3064.00	7527.00	47.00
2013-01-05	6458.00	10699.00	57.00	4852.00	10550.00	47.00

Dati dal 01/2013 al 08/2017

Obiettivo: prevedere 28 giorni di vendite per le sei serie storiche.

02

1. Ottenere un **valore medio del MAPE minore** rispetto al modello base.

2. Il **MAPE** deve essere **stabile**: cioè i **modelli** di **previsione** sono **robusti** rispetto ai nuovi dati.

Darts

Serie storica: definizione e componenti

03

Insieme temporalmente **ordinato** di **osservazioni**, in base ad un intervallo di tempo **regolare**, detto **frequenza**: ogni giorno, minuto, mese, secondo ecc.

La serie storica è una **sequenza temporale**.

La dimensione del **tempo crea** una **dipendenza** tra le osservazioni.

Componenti:

- **Livello**
- **Trend**
- **Stagionalità**
- **Rumore**

Trend e stazionarietà

04

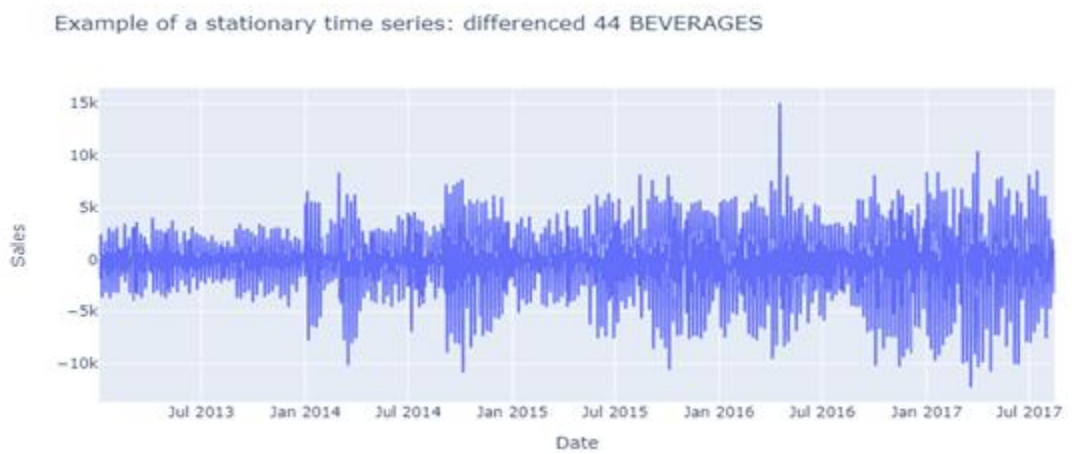
Il **trend** è l'**andamento di lungo periodo**,

cioè la **tendenza**: si individua come differenza tra il dato iniziale e quello finale della serie.

La **stazionarietà** è la proprietà per cui le **statistiche descrittive non variano** tra spezzoni casuali della serie.

La **media** e la **deviazione standard** rimangono pressoché **invariate**.

Condizione necessaria per ottimizzare le performance di **modelli ARIMA e LSTM**.



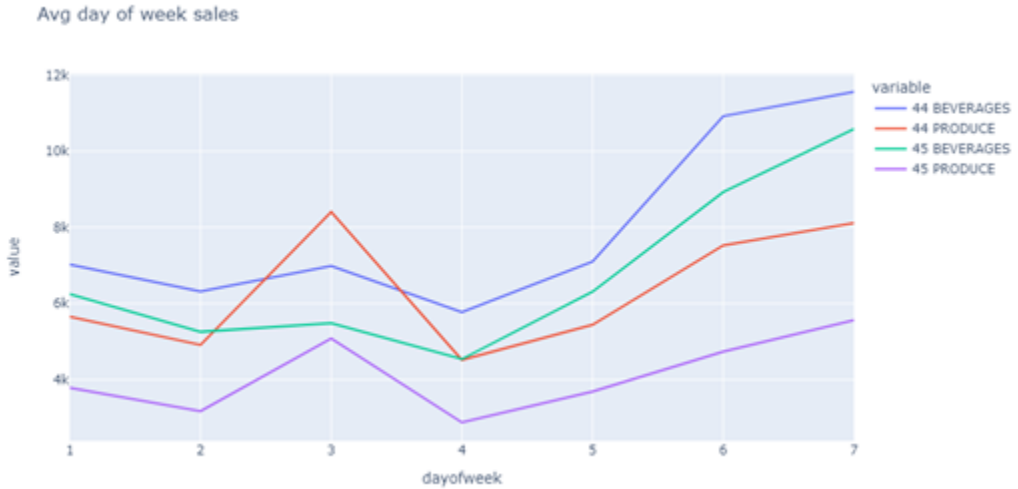
Stagionalità e covariate future

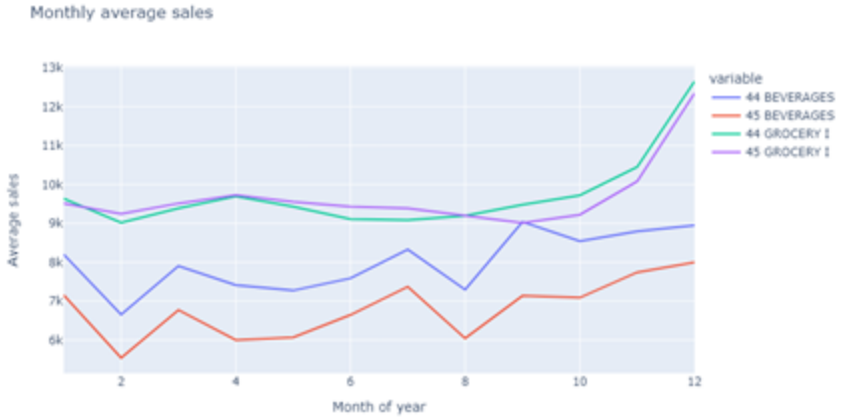
04

Le sei serie presentano **stagionalità settimanale** e **stagionalità annuale**.

La stagionalità **settimanale** è la **principale**, quella maggiormente **evidente**;

la stagionalità **annuale** è utile per **identificare** i periodi di **maggiore attività**: soprattutto il **periodo natalizio**.





Processo di previsione

06

- Valutare necessità **serie differenziata** e **dati scalati**.
- Individuazione dei **lag**, cioè i **predittori** principali, tramite funzione di **AUTOCORRELAZIONE PARZIALE**.
- Uso **covariate temporali** dei **giorni della settimana** e **settimane dell'anno.**
- Individuazione **miglior set di parametri** per ogni modello.

Modello base

Mape 13,03%

Modello naive seasonal k=7.

Modelli per serie stazionarie

Mape 9,13%

Mape 10,72%

Modelli AR e LSTM.

Modelli per serie non stazionarie

Mape 7,62%

Accuracy +42%

Mape 10,41%

Modelli Random forest e Boosting.

Ensemble model finale e backtesting

07

Il **modello** che **performa mediamente meglio** è il **Random forest**.

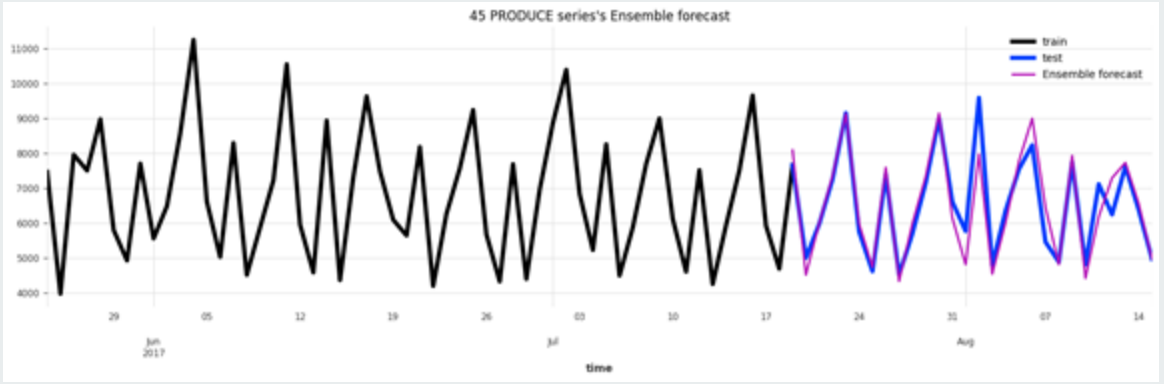
Il backtesting consente di **stabilire** quali **modelli includere nell'ensemble model** e con quale peso.

Il **modello ensemble** è una **combinazione lineare principalmente dei modelli Random forest** e **AR**.

	Model	Mean
0	Random Forest mape	7.62
1	ensemble mape	7.89
2	Arima mape scaled data and weeks	9.13
3	auto arima mape	9.45
4	lightgbm mape	10.41
5	Bagging mape	10.67
6	LSTM mape	10.72
7	Arima mape cov days	10.82
8	Arima mape scaled data	11.05
9	Random forest diff and weeks mape	12.80
10	Baseline mape	13.03
11	LSTM no val mape	19.22

Conclusioni

09



Previsioni attendibili generano valore per l'impresa = **Redditività maggiore**.

Ottimizzazione livelli magazzino e **tempistiche** di **approvvigionamento** = **meno costi**.

Prodotti disponibili nei **tempi** e nei **punti vendita corretti** e proprio nelle **quantità richieste** dai clienti.

Clienti soddisfatti = clienti di fiducia **= ricavi garantiti**.