Sales Forecasting for Promotional Flyers

Agenda

1. Introduction
2. Data Preprocessing
3. Feature Engineering
4. Model exploration
5. Final approach
6. Discussions

1 Introduction

Our goal & datasets

Predict the sales units for each product featured in promotional flyers from November to December 2023.

Anagrafica_volantini, Risultati_prodotti_volantini, Storico_quantità Gfk_caratteristiche, KPI_prodotti volantini. Icecat_caratteristche

2 Preprocessing

Data manipulation

Prezzo_listino

Recovered the missing prices by averaging the prices of products with the same art_code, for the remaining NaNs, we used historical revenue divided by historical quantity.

Prezzo_promo

For the missing values, it was calculated using the percentage discount and the list price.

Stock_pz

All negative values have been replaced with 0.

Sconto_perc

Measured the absolute value for the negative ones, while multiplying by 100 all those less than one. For the missing values, we calculated the discount using the promotional price and the list price.

Focus on gfk_caratteristiche

Initially, we did not know how to use this dataset because it contained a large number of categorical variables.

We noticed that for the product group, the type of characteristic was similar within the same column.

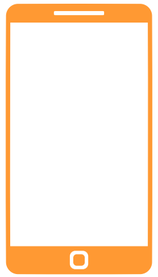
We then split it by product group and performed ordinal encoding, avoiding the creation of a large number of variables.

Final DataSet

Anagrafica volantini, **Risultati prodotti volantini**, **Storico quanittà**, **Gfk caratteristiche** and **Kpi prodotti volantini** got merged together on “**art_cod”** and “**codice volatino**” and then split for **product group**

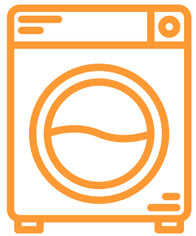
Resulting in 5 different datasets, one for each product group:

Smartphone



PC

Wash



TV

Core_wear

3

Feature engineering

New variables

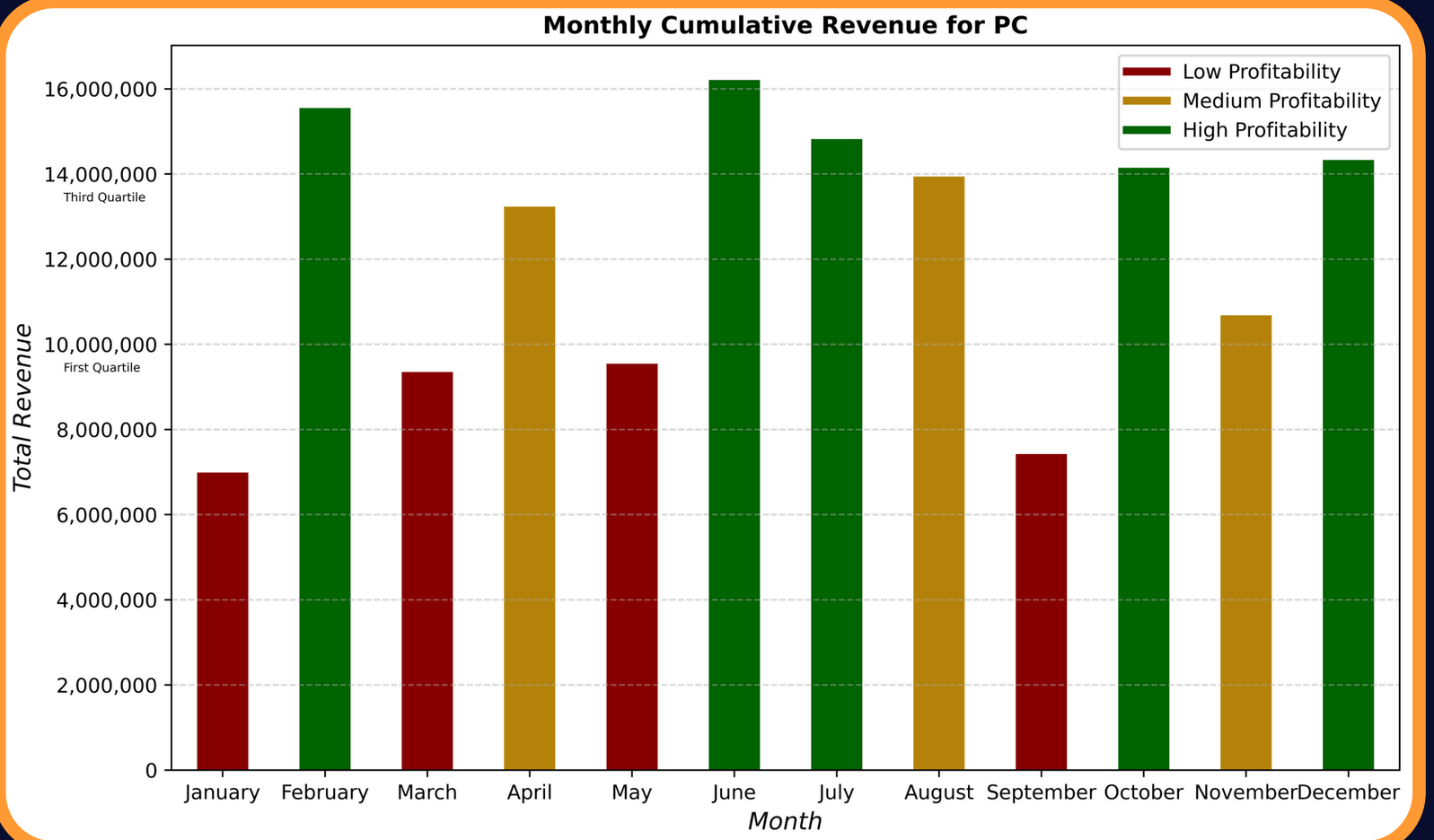
Flyer duration in days

Average discount by campaign name

Average Monthly Revenue

Average discount per flyer

Average Monthly Revenue



1. Computed as the **monthly cumulative revenue** of each product group

2. Then groupped the months into **profitability classes**: low, medium, and high.

Average Discount by Campaign Name

1. We grouped the **campaign names** into different macro areas (e.g., all campaigns regarding ***Black Friday***, ***Scontissimi,*** ***Tasso 0***, etc.)

2. Computed the **average discount** related to each group of campaigns

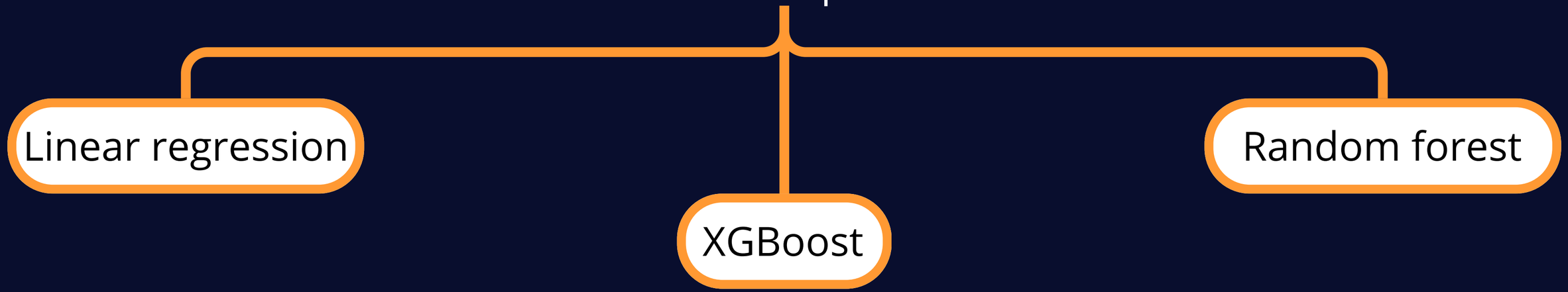
3. Finally, the column ***“Campaign Name”*** was dropped

4

Model exploration

Our approach

We decided to implement 3 different models and a benchmark to evaluate their performance



Our benchmark consists consists of multiplying the average daily sales per product by the days the flyer is online

Work phases

The work can be divided into 3 different phases.
These phases differs from one another based on the train and test split.

1

5-folds cross validation using **the entire dataset**

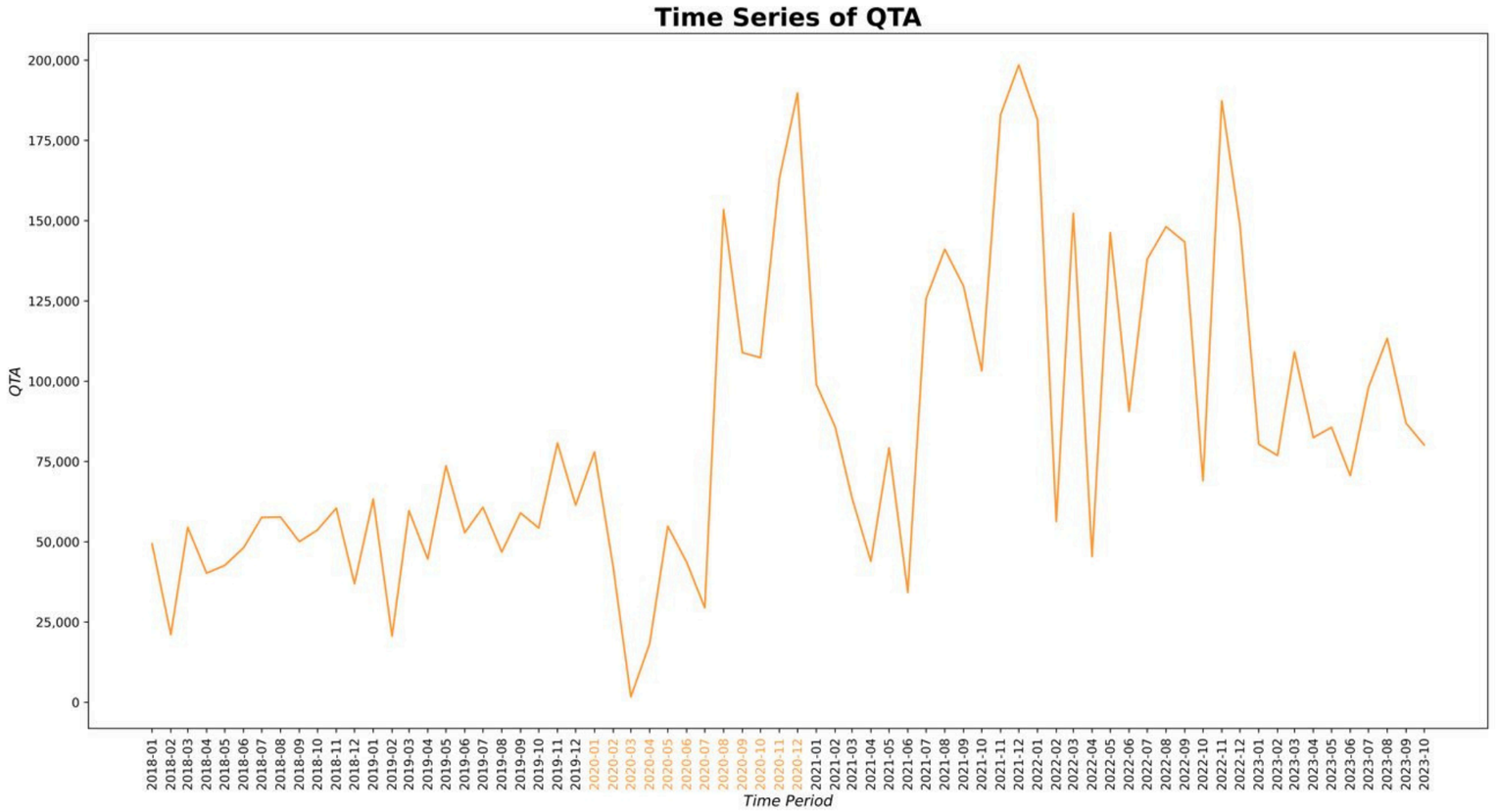
2

5-folds cross validation **excluding 2020**

3

The **entire dataset** was included in the training set, **except data from 2020**, and used as test set **August, September and October 2023**

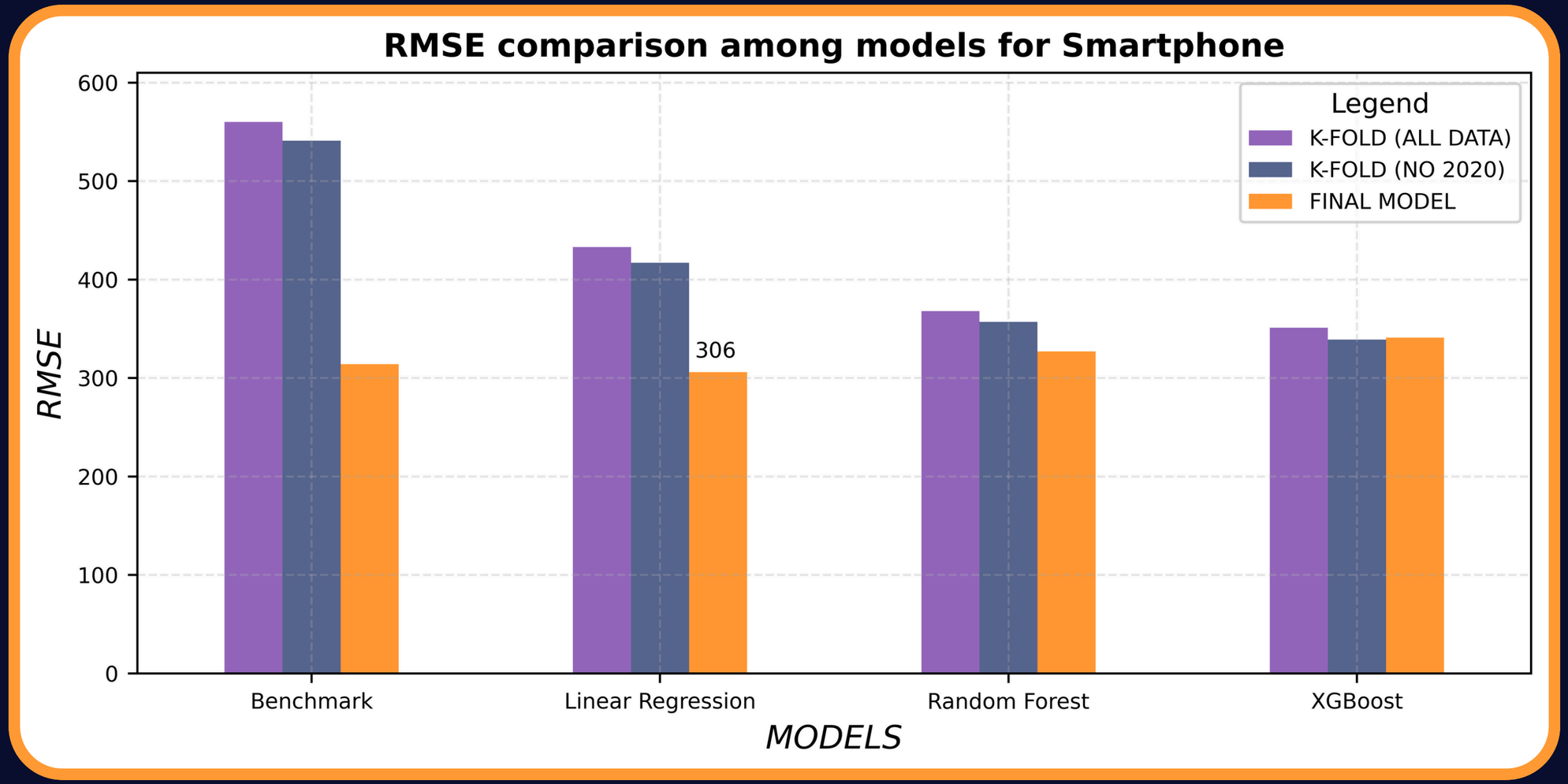
What about 2020?



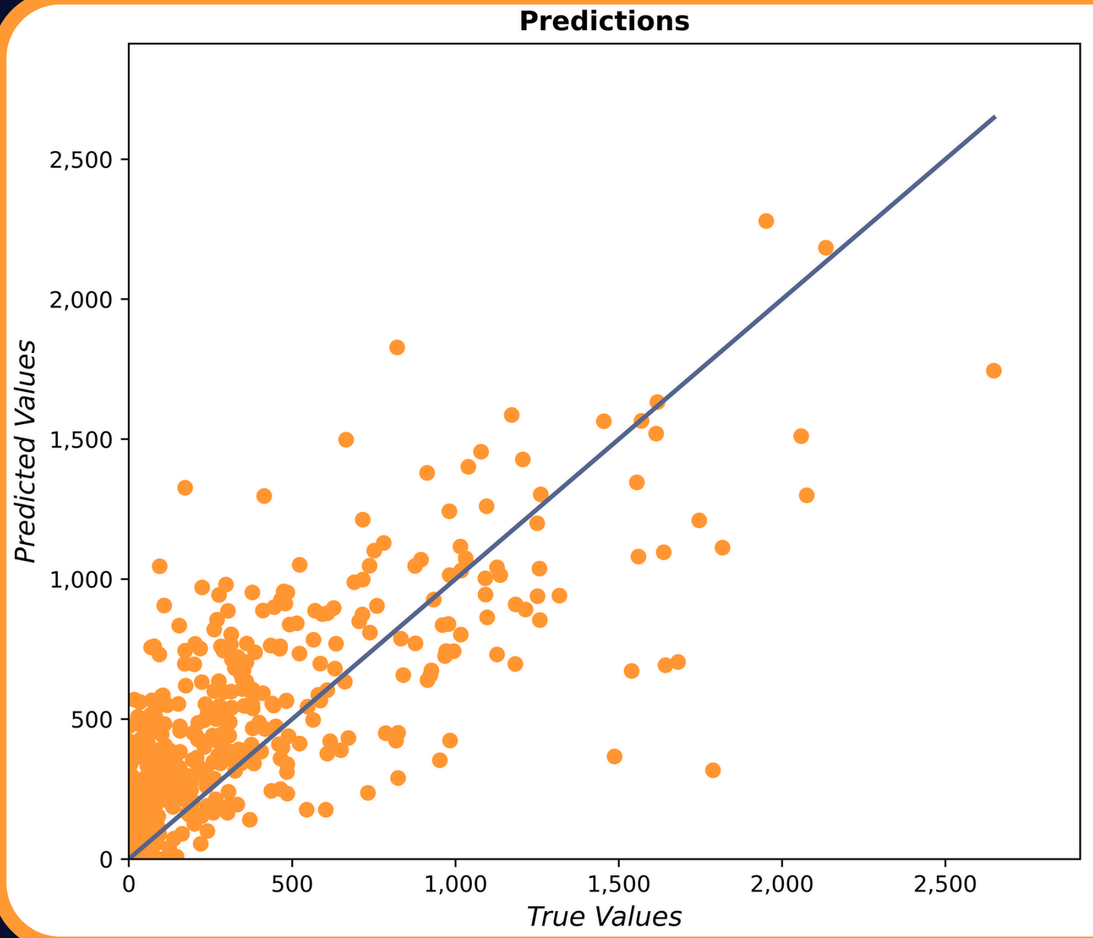
The year **2020** was **removed** from our analysis under the assumption that since it was the period of the **COVID-19** pandemic, sales would not follow a normal trend, thus **impacting the performance of our models**.
This particular trend can also be seen from the graph above.

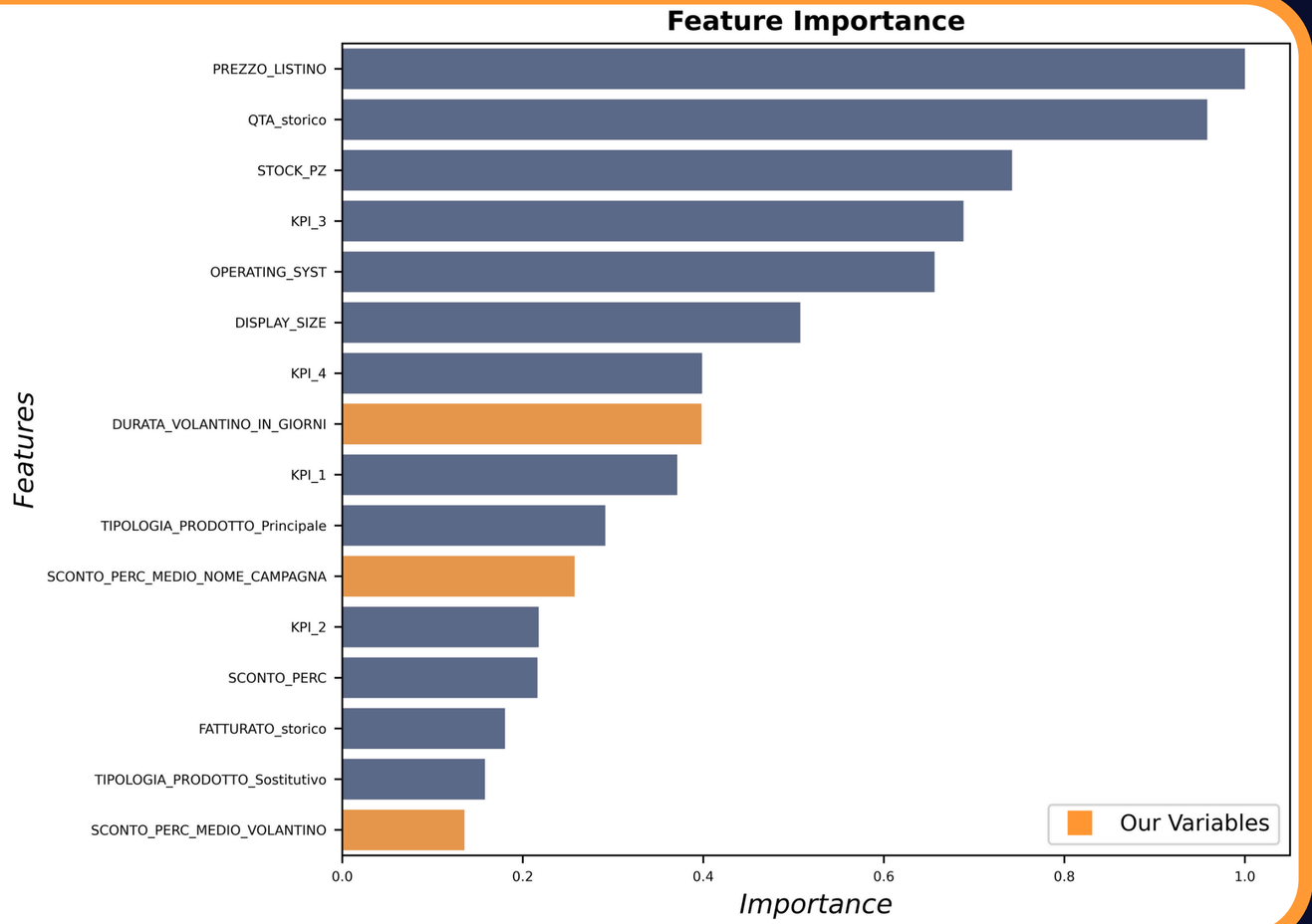
5 Final models

Smartphone

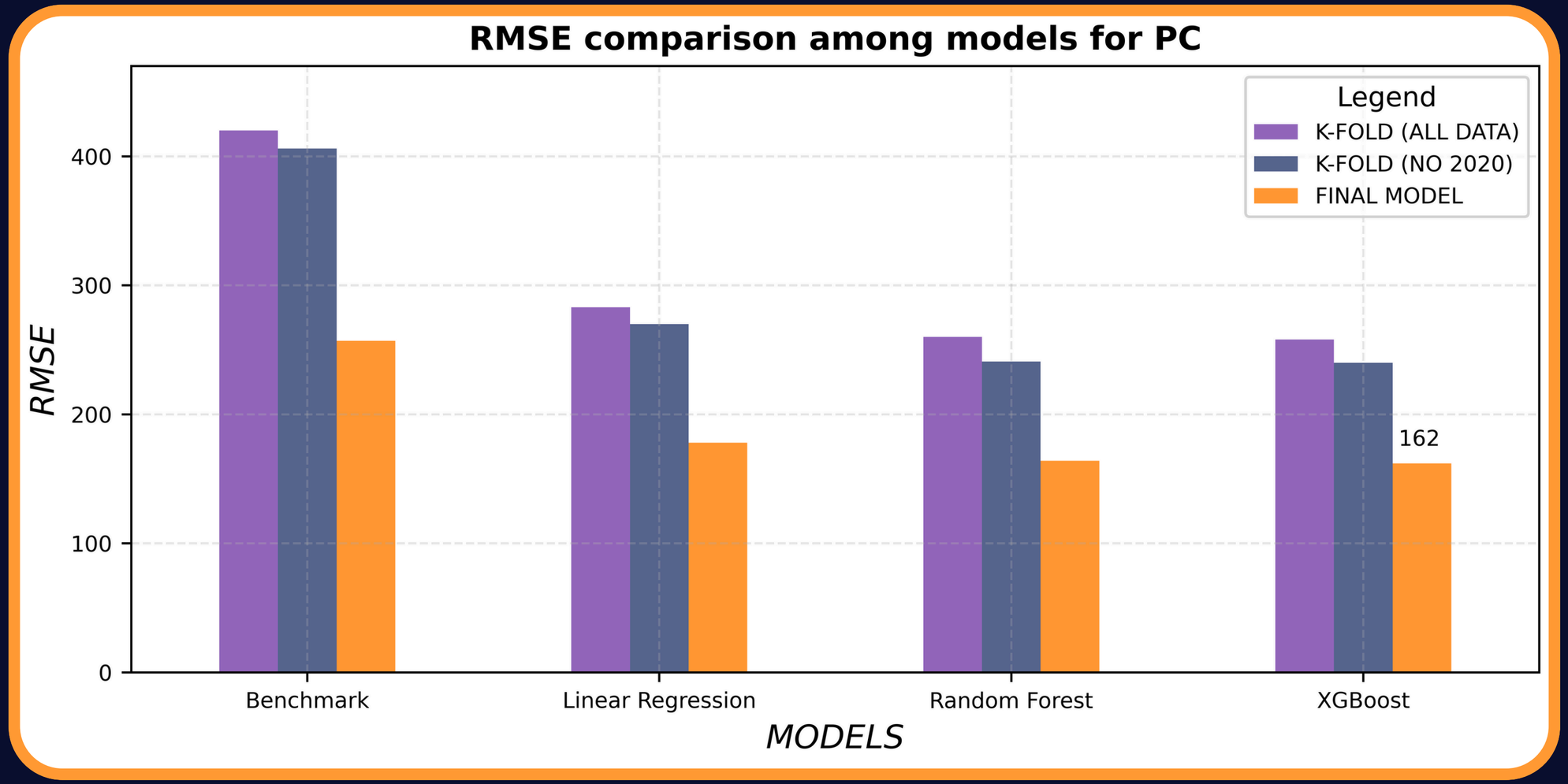

RMSE comparison among models for Smartphone
Legend
K-FOLD (ALL DATA)
K-FOLD (NO 2020)
FINAL MODEL
RMSE
MODELS
600
500
400
300
200
100
0
306
Benchmark
Linear Regression
Random Forest
XGBoost

Smartphone - Linear Regression

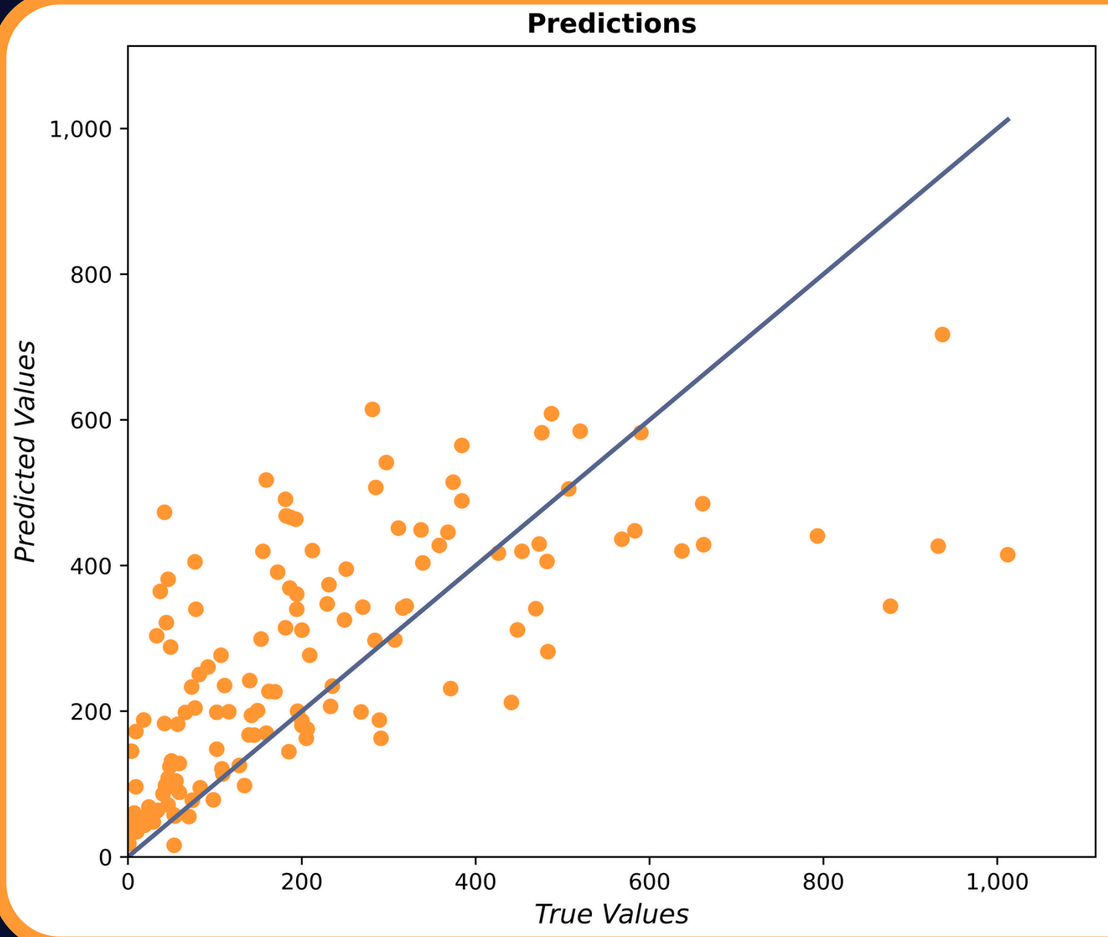


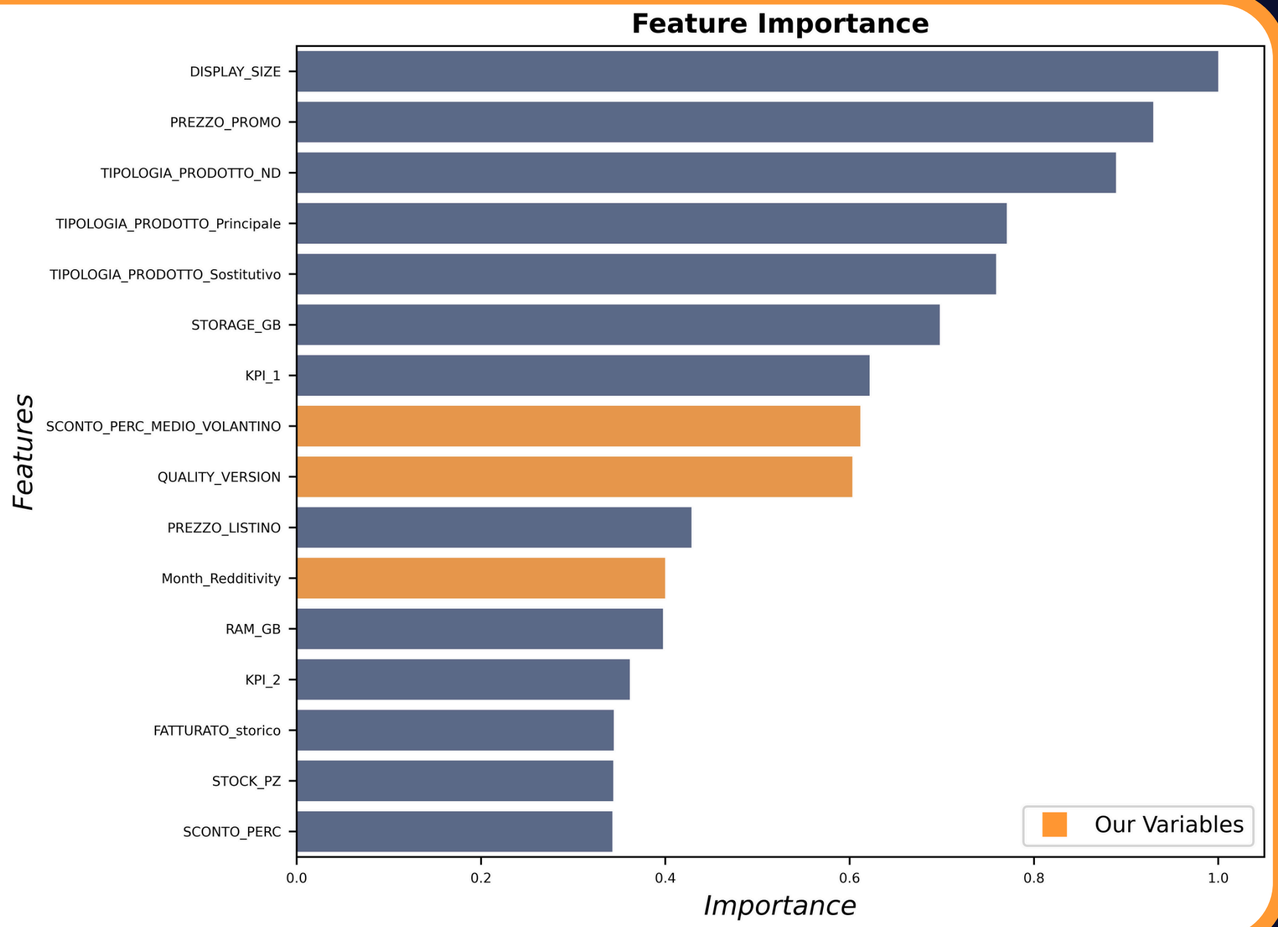


PC

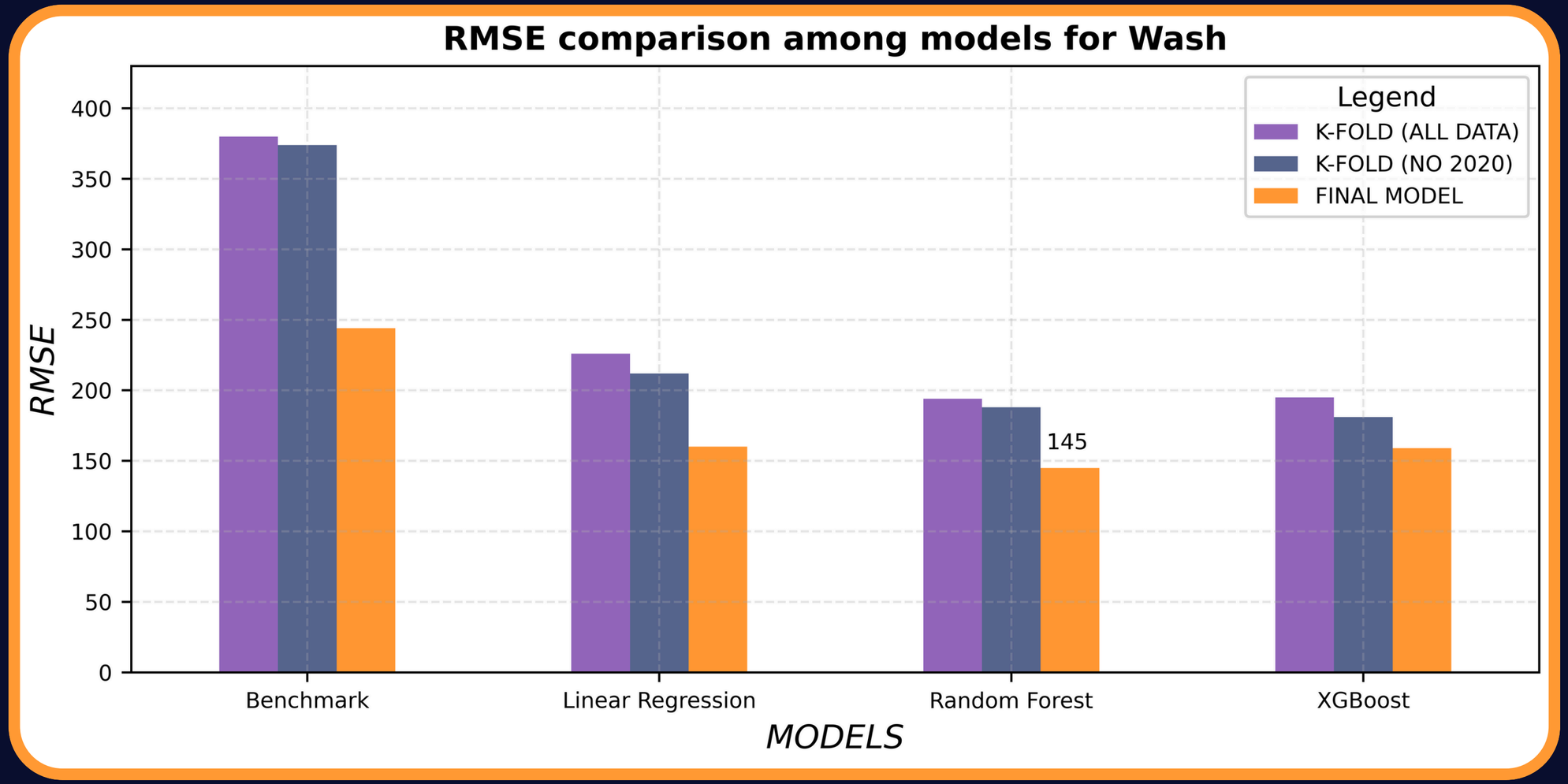

RMSE comparison among models for PC
Legend
K-FOLD (ALL DATA)
K-FOLD (NO 2020)
FINAL MODEL
RMSE
400
300
200
100
0
162
Benchmark
Linear Regression
Random Forest
XGBoost
MODELS

PC - XGBoost

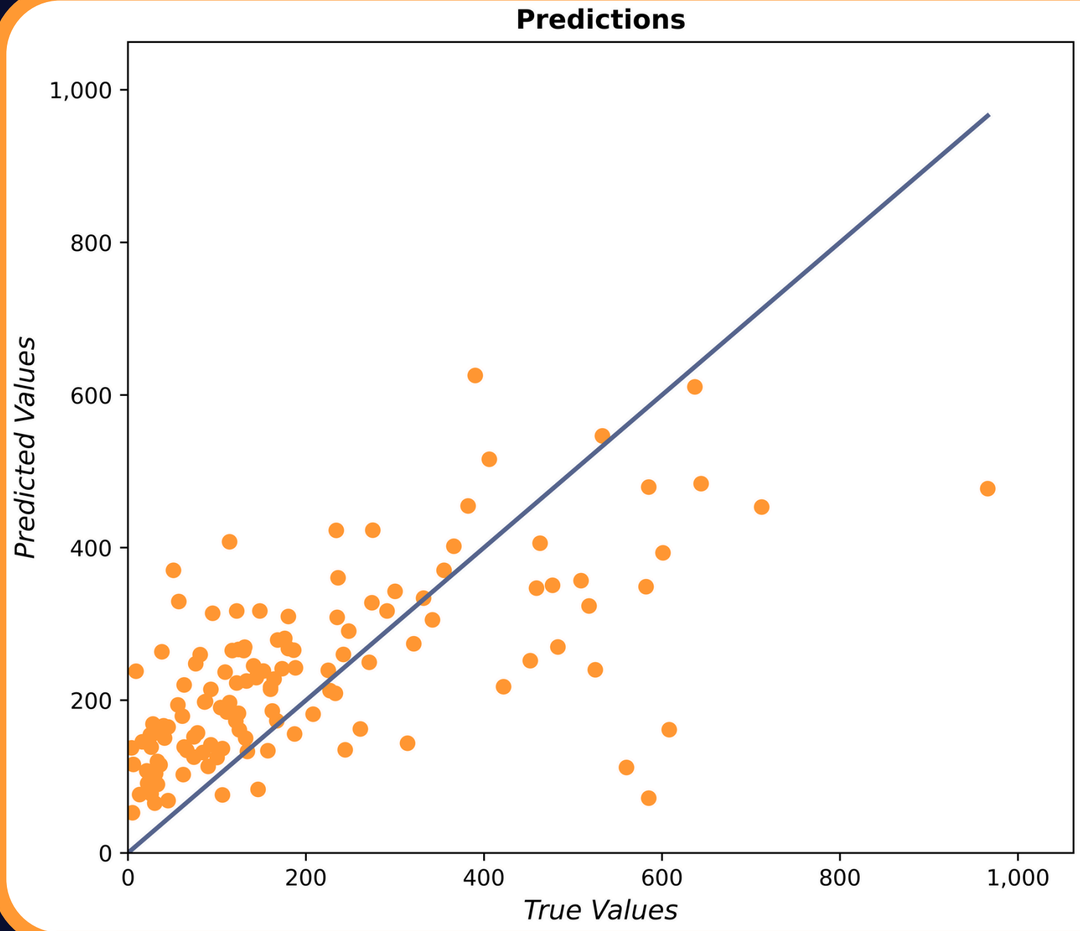

Predictions
Predicted Values
True Values
0
200
400
600
800
1,000

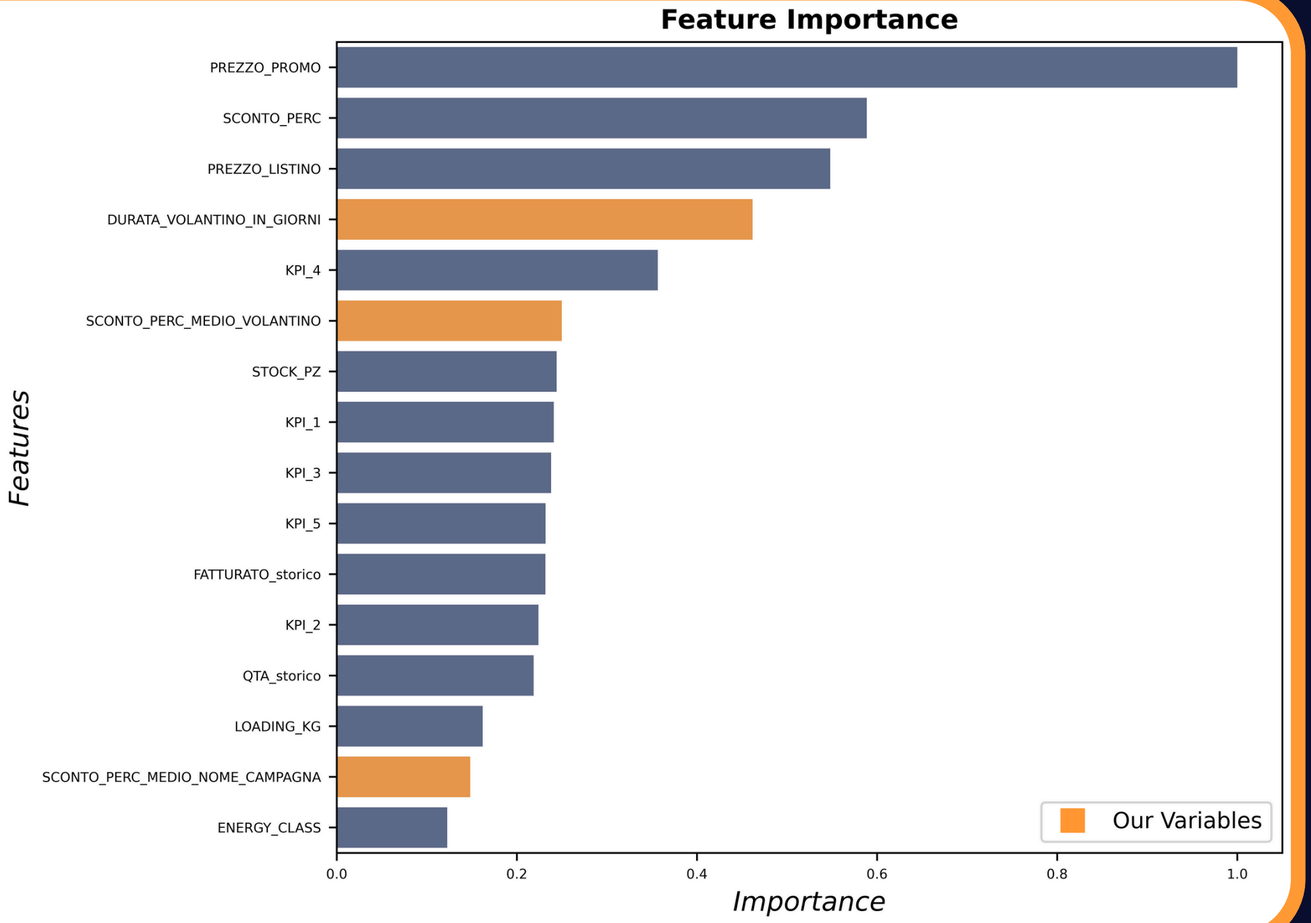

Feature Importance
Features
Importance
DISPLAY_SIZE
PREZZO_PROMO
TIPOLOGIA_PRODOTTO_ND
TIPOLOGIA_PRODOTTO_Principale
TIPOLOGIA_PRODOTTO_Sostitutivo
STORAGE_GB
KPI_1
SCONTO_PERC_MEDIO_VOLANTINO
QUALITY_VERSION
PREZZO_LISTINO
Month_Redditivity
RAM_GB
KPI_2
FATTURATO_storico
STOCK_PZ
SCONTO_PERC
Our Variables
0.0
0.2
0.4
0.6
0.8
1.0

Wash

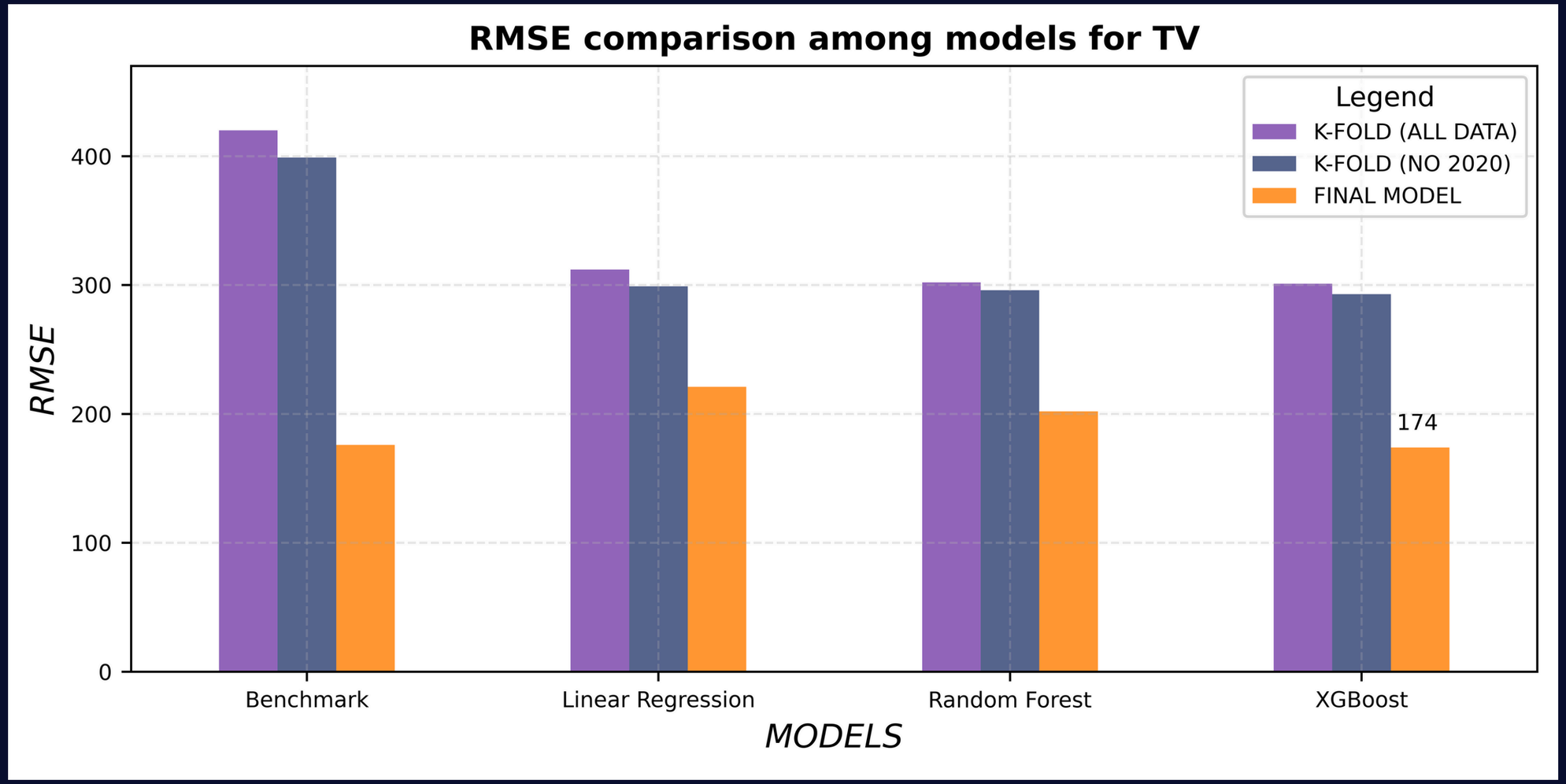

RMSE comparison among models for Wash
Legend
K-FOLD (ALL DATA)
K-FOLD (NO 2020)
FINAL MODEL
RMSE
400
350
300
250
200
150
100
50
0
145
Benchmark
Linear Regression
Random Forest
XGBoost
MODELS

Wash - Random Forest

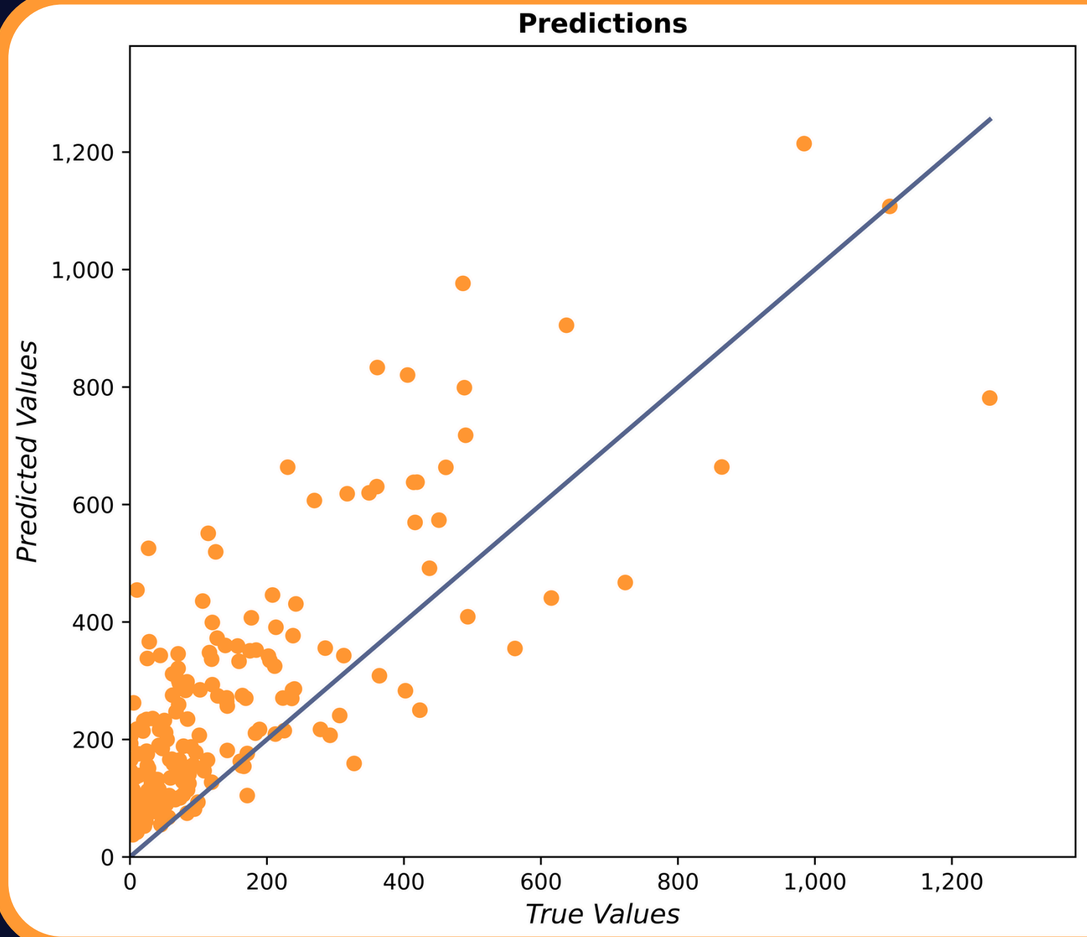


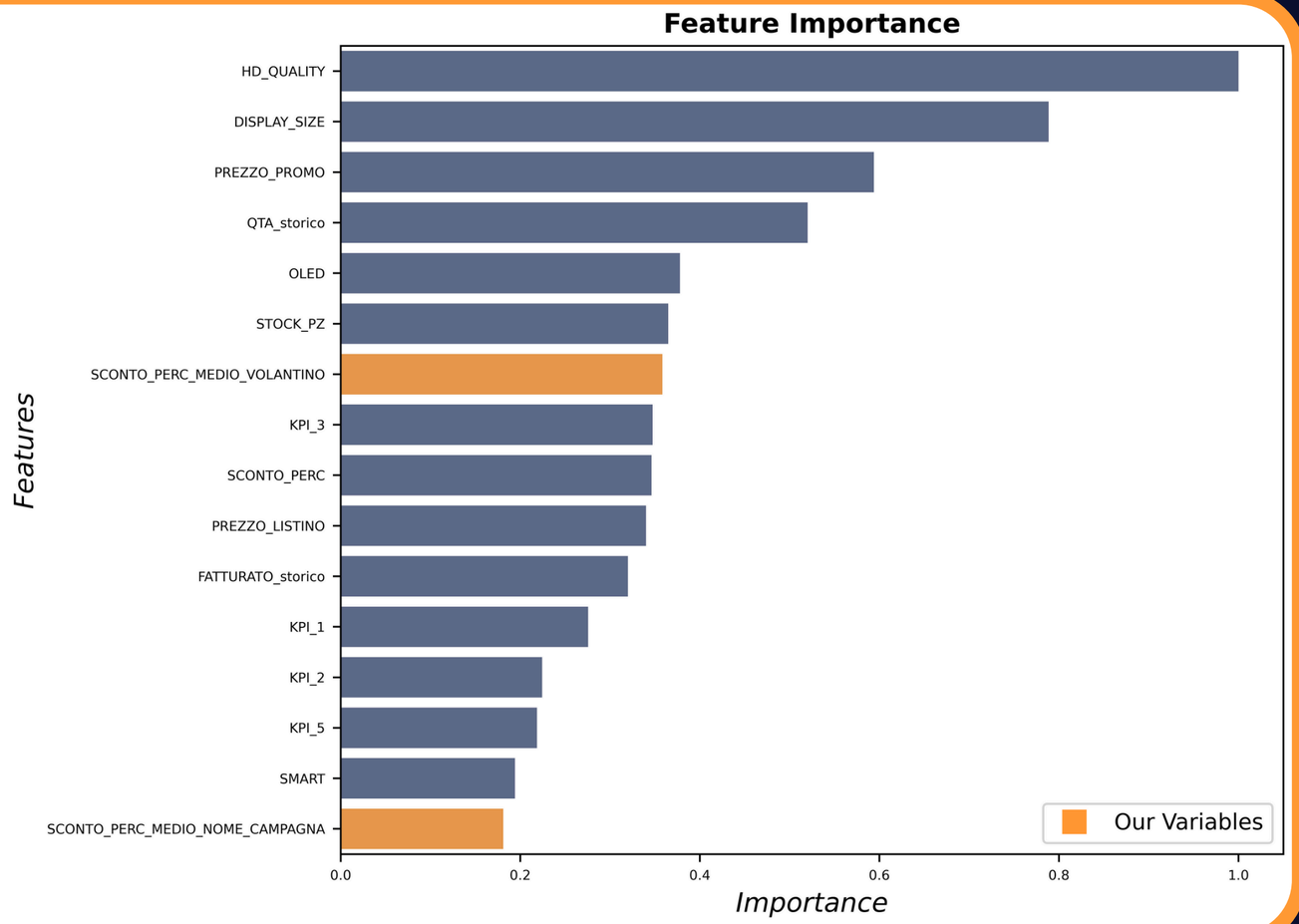


TV

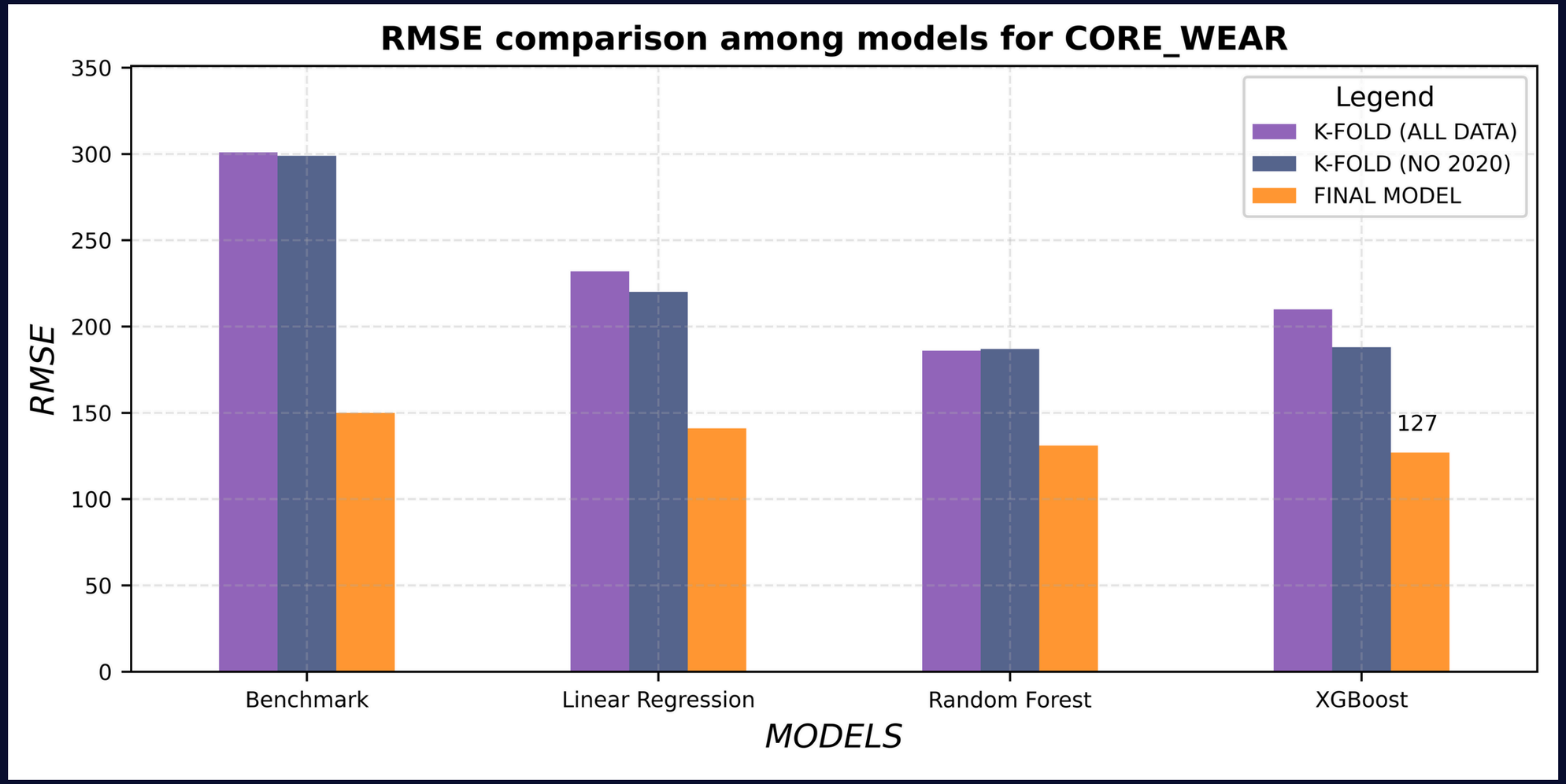

RMSE comparison among models for TV
Legend
K-FOLD (ALL DATA)
K-FOLD (NO 2020)
FINAL MODEL
RMSE
0
100
200
300
400
174
Benchmark
Linear Regression
Random Forest
XGBoost
MODELS

TV - XGBoost

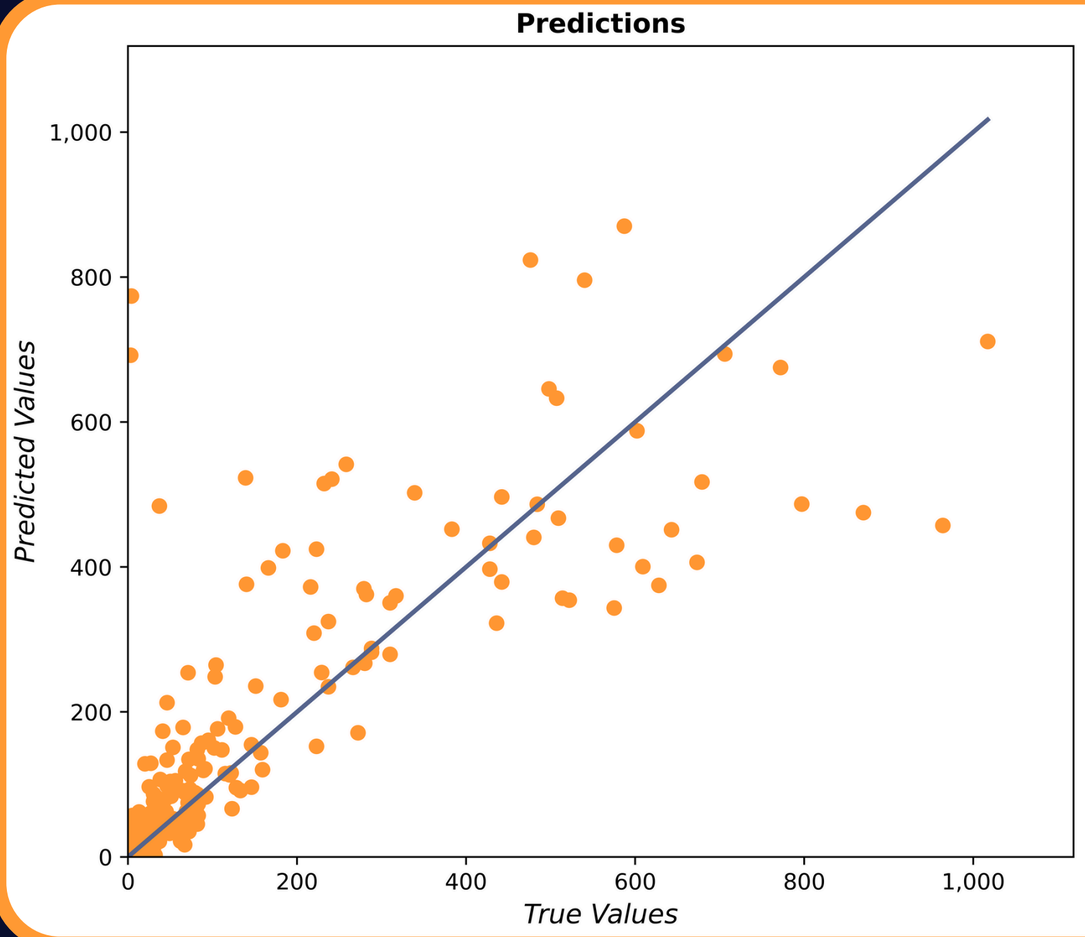

Predictions
Predicted Values
True Values
0
200
400
600
800
1,000
1,200

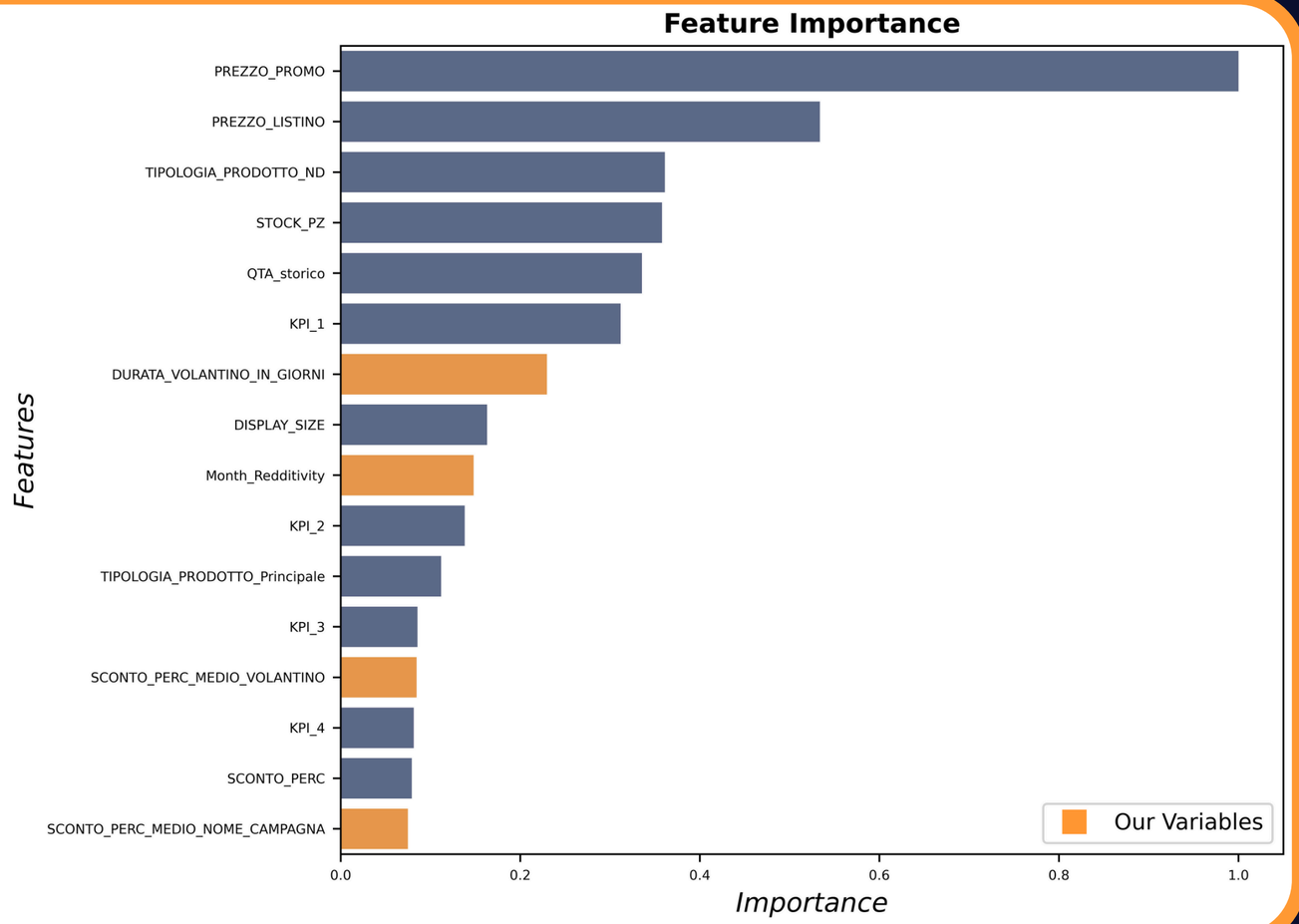

Feature Importance
Features
Importance
HD_QUALITY
DISPLAY_SIZE
PREZZO_PROMO
QTA_storico
OLED
STOCK_PZ
SCONTO_PERC_MEDIO_VOLANTINO
KPI_3
SCONTO_PERC
PREZZO_LISTINO
FATTURATO_storico
KPI_1
KPI_2
KPI_5
SMART
SCONTO_PERC_MEDIO_NOME_CAMPAGNA
Our Variables
0.0
0.2
0.4
0.6
0.8
1.0

Core Wearables



Core Wearables - XGBoost





6 Discussion

Discussion

Preprocessing;

Feature engineering;

Computational time;

Good prediction for core wear, wash, TV and PC;

Hyperparametertuning;

Implementation of new variables;

Poor prediction for smartphone;

Dimensionality reduction;

Thank You!

for the Attention

Vincenzo Camerlengo	773731
Riccardo Paolantoni	773691
Raffaele Torelli	775831