Report: BIP project

Alberto Parravicini, Simone Ripamonti, Luca Stornaiuolo

DATA PREPROCESSING

- INITIAL ATTRIBUTES

    - "Zona": [STRING];
    - "Area": [STRING];
    - "Sottoarea": [STRING];
    - "Categoria_prodotto": [STRING];
    - "Data": [STRING];
    - "Vendite": [NUMBER];
    - "LATITUDINE": [NUMBER];
    - "LONGITUDINE": [NUMBER];

- FINAL ATTRIBUTES

    - Attributes extracted from "Categoria_prodotto"

        * "prod": [NUMBER] {string manipulation; cast to integer};

    - Attributes extracted from "Zona", "Area" and "Sottoarea":

        * "zona": [NUMBER] {string manipulation; cast to integer};
        * "area": [NUMBER] {string manipulation; cast to integer};
        * "sottoarea": [NUMBER] {string manipulation; cast to integer};
        * "key": [STRING] {string manipulation};

    - Attributes extracted from "Data"

        * "data": [STRING];
        * "giorno_mese": [NUMBER] {cast to date; date manipulation};
        * "giorno_settimana": [NUMBER] {cast to date; date manipulation};
        * "giorno_anno": [NUMBER] {cast to date; date manipulation};
        * "mese": [NUMBER] {cast to date; date manipulation};
        * "settimana_anno": [NUMBER] {cast to date; date manipulation};
        * "anno": [NUMBER] {cast to date; date manipulation};
        * "weekend": [NUMBER (BOOLEAN)] {cast to date; date manipulation};
        * "stagione": [STRING] {cast to date; date manipulation};
        * "primo_del_mese": [NUMBER (BOOLEAN)] {cast to date; date manipulation};

    - Attributes extracted from "Data" and GPS information

        * "vacanza": [NUMBER (BOOLEAN)] {cast to date; date manipulation};

    - Attributes extracted from "Vendite"

        * "vendite": [NUMBER] {attribute renaming};
        * "sp_anno": [NUMBER] {number manipulation};
        * "sp_mese": [NUMBER] {number manipulation};
        * "sp_settimana": [NUMBER] {number manipulation};
        * "vendite_giorn_tot": [NUMBER] {number manipulation};
        * "vendite_missing": [NUMBER (BOOLEAN)] {number manipulation};
        * "vendite_giorn_prod": [NUMBER] {number manipulation; PREDICTION PROCESS};

    - Attributes extracted from GPS information

        * "latitudine": [NUMBER] {attribute renaming};
        * "longitudine": [NUMBER] {attribute renaming};

* "cluster3": [NUMBER] (GPS data manipulation; CLUSTERING PROCESS);
* "cluster6": [NUMBER] (GPS data manipulation; CLUSTERING PROCESS);
* "cluster20": [NUMBER] (GPS data manipulation; CLUSTERING PROCESS);

CLUSTERING PROCESS

- HIERARCHICAL CLUSTERING (GPS information)

  – Elbow Method Analysis: number of cluster = 6;

- KMEANS CLUSTERING (GPS information)

  – "cluster6": from Elbow Method Analysis;
  – "cluster3": NORD, CENTRO, SUD Analysis;
  – "cluster20": REGIONI Analysis;

PREDICTION PROCESS

- RANDOM FOREST

  – Single model for all sottoarea;
  – Ranger random forest model builder;
  – Ranger random forest model builder;

- SARIMA PREDICTION

  – Time Series Casting;
  – Correlation Analysis;
  – Regressors (External Input);

- XGBOOST PREDICTION

  – Single XGBoost;
  – Multiple XGBoost;

MODELS EVALUATION

- SSE;
- MAPE;