Modello per la previsione del valore economico degli acquisti

Saverio Pulizzi,
Candidatura Posizione Data Scientist @Team Digitale

Agenda

- Quadro generale del processo di acquisto di beni e/o servizi
- Assunzione base
- La domanda di ricerca
- Preparazione dataset
- Analisi esplorativa
- Il Modello
- Lezioni apprese ed opportunità
- Considerazioni

Quadro generale sul processo di acquisto di beni e/o servizi

Il processo di acquisto

- Al momento esistono **4 distinti canali** che permettono l'approvvigionamento di beni e/o servizi ai vari enti della PA: accordi quadro, mercato elettronico PA, convenzioni e sistema dinamico.

- Ogni canale permette una tipologia di acquisto differente sotto forma di **ordine diretto** o di **negoziazione**

GLI STRUMENTI DI ACQUISTO

Gli strumenti di acquisto / vendita, attraverso cui le imprese offrono i propri beni e servizi alla P.A. e le Amministrazioni effettuano acquisti, sono le Convenzioni, gli Accordi quadro, il Mercato Elettronico e il Sistema dinamico di acquisizione.

CONVENZIONI

Le Convenzioni sono dei contratti che le Amministrazioni possono utilizzare per l'acquisto o il noleggio di beni e servizi.

Approfondisci >>

Cerca l'etichetta CONVENZIONI

ACCORDI QUADRO

Gli Accordi quadro sono contratti quadro aggiudicati da Consip a uno o più fornitori che le P.A. possono utilizzare per acquistare prodotti e servizi.

Approfondisci >>

Cerca l'etichetta ACCORDI QUADRO

MERCATO ELETTRONICO

È il mercato digitale per gli acquisti sotto soglia comunitaria di beni, servizi e lavori di manutenzione.

Approfondisci >>

Cerca l'etichetta MePA

SISTEMA DINAMICO

È un mercato digitale per gli acquisti di beni e servizi dove le imprese richiedono l'ammissione ai bandi attivi e le P.A. pubblicano i propri Appalti specifici.

Approfondisci >>

Cerca l'etichetta SDA

Amministrazioni

PA registrate al programma di razionalizzazione degli acquisti.

- Dal 2018 ad oggi: 26 diverse tipologie amministrative includenti 23,265 enti partecipanti

- Gli enti sono dislocati in 20 regioni, 111 provincie e 8,112 comuni

Fornitori

Imprese partecipanti al programma di razionalizzazione degli acquisti.

- Dal 2018 ad oggi:157,020 imprese partecipanti suddivise in 20 forme societarie diverse
- 798 distinte denominazioni di lotto
- 1.57 Milioni di transazioni

Cataloghi Regionali

Elenco dei beni e servizi offerti dai fornitori abilitati al Mepa o aggiudicatari di Convenzioni e Accordi Quadro

- Dal 2018 ad oggi: 3.5 Milioni di beni e 470 Mila servizi registrati
- 1,622 tipologie distinte di prodotti/servizi offerti
- I tre prodotti/servizi più comunemente disponibili tra i cataloghi sono:
    - servizi di formazione
    - servizi di informazione comunicazione e marketing
    - ferramenta
- Software di gestione ed accessori per alimentazione risultano essere tra i primi beni più comunemente disponibili
- Più del 99% di questi beni/servizi sono disponibili tramite MePa

Cataloghi del Programma

Elenco dei beni, servizi e lavori disponibili tramite le iniziative del programma per la razionalizzazione degli acquisti.

- 1,878 prodotti/servizi diversi offerti appartenenti a 15 categorie merceologiche
- I tre prodotti/servizi più comunemente disponibili tra i cataloghi sono:
    - Servizi opzionali per autoveicoli
    - Multi servizio integrato energia
    - Energia elettrica
- Gli strumenti di acquisto per questa tipologia di beni/servizi sono:
    - MePa ~36%
    - Convenzione ~35%
    - Accordo quadro ~9%
    - SDA ~20%

Bandi & Gare

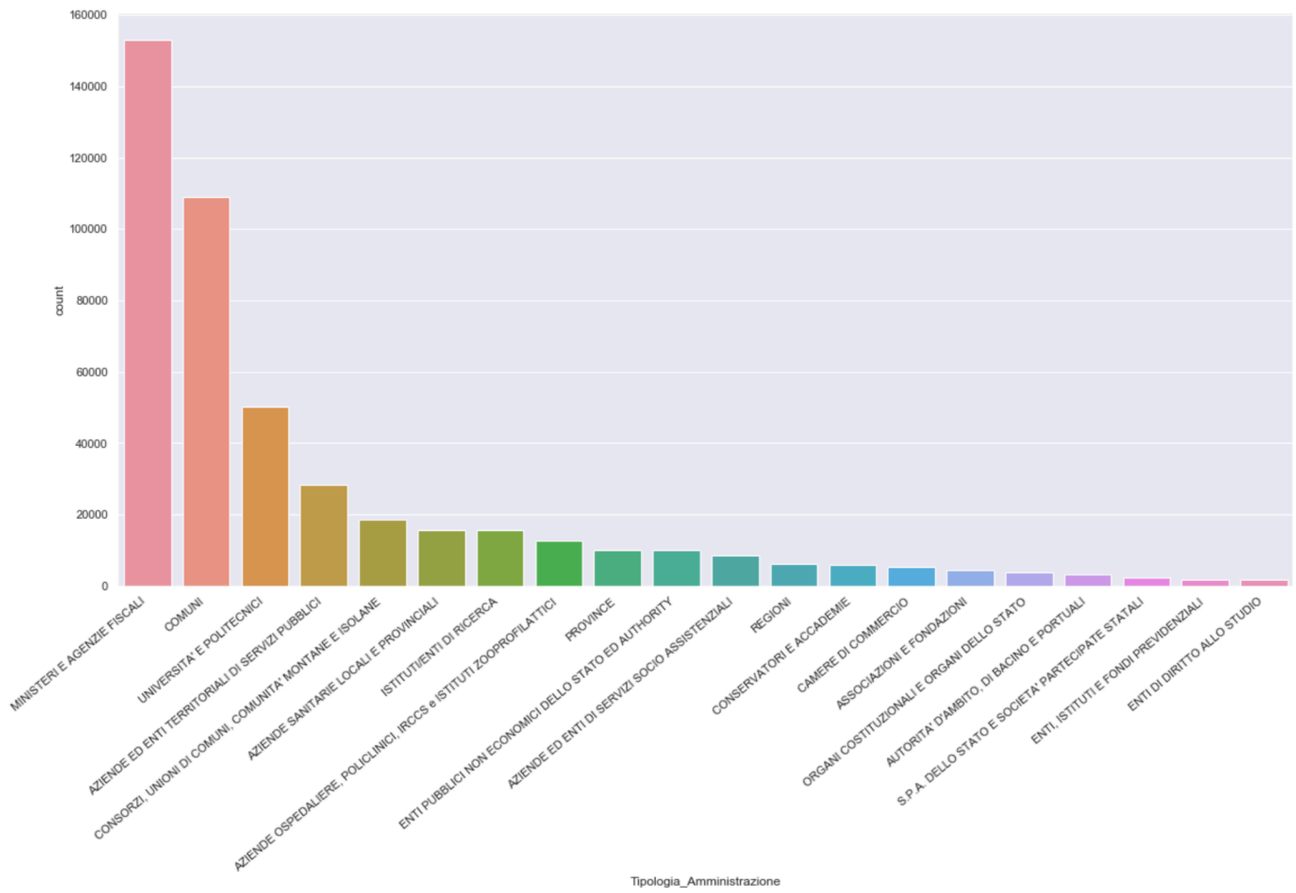
Iniziative di gara e lotti pubblicati nell'ambito del programma di razionalizzazione degli acquisti.

- Dal 2018 ad oggi: 1,097 bandi unici svolti
- 1,178 Lotti, 3 tipologie di lotto con 14 diverse categorie merceologiche
- Un giro di affari totale di 144 miliardi

Acquisti & Negoziazioni (1)

Acquisti e negoziazioni effettuate dalle PA attraverso le varie piattaforme telematiche

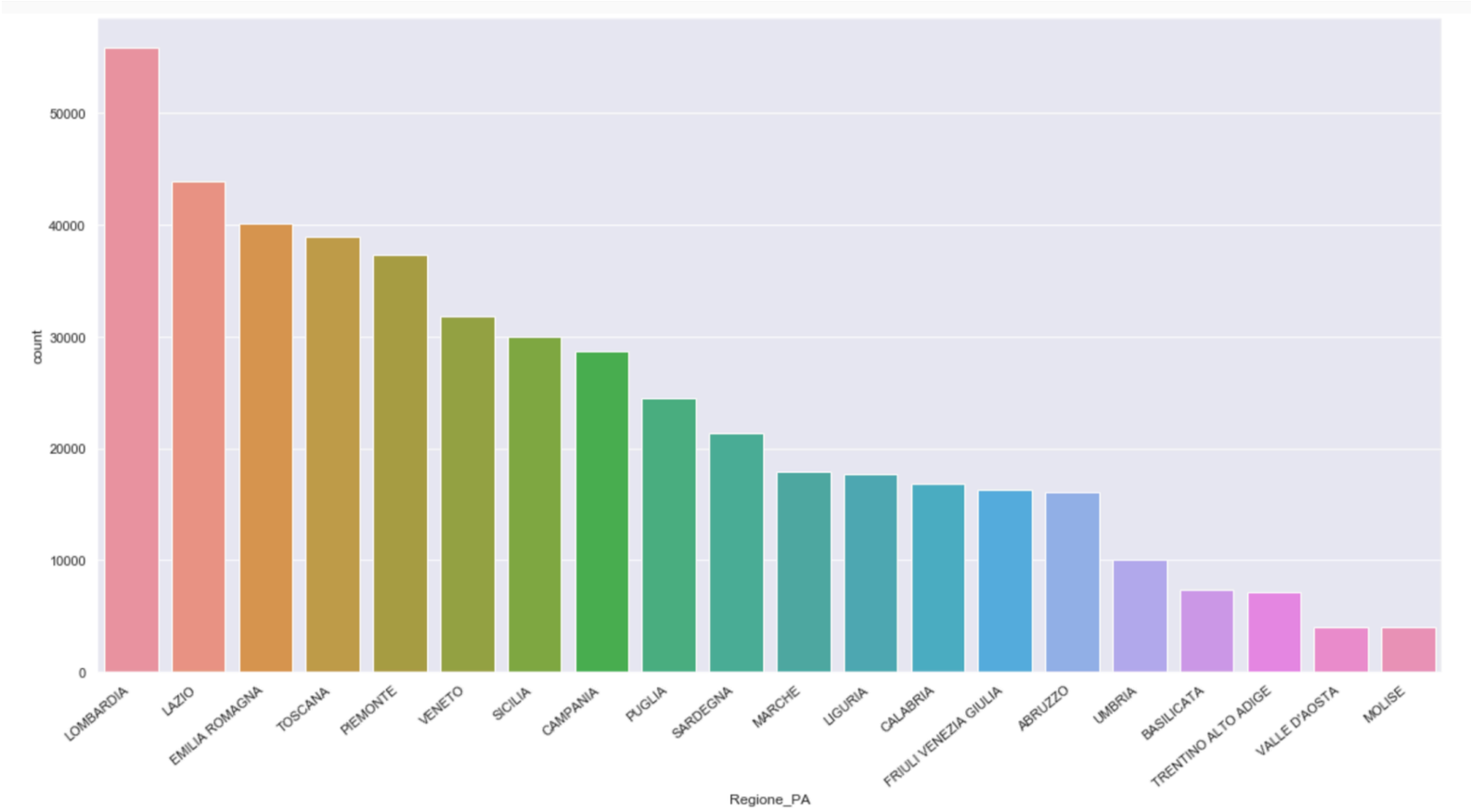
Ministeri ed agenzie fiscali insieme ai **comuni** sono gli enti amministrativi che hanno registrato il maggior numero di acquisti e/o negoziazioni dal 2018 ad oggi.



Acquisti & Negoziazioni (2)

Acquisti e negoziazioni effettuate dalle PA attraverso le varie piattaforme telematiche

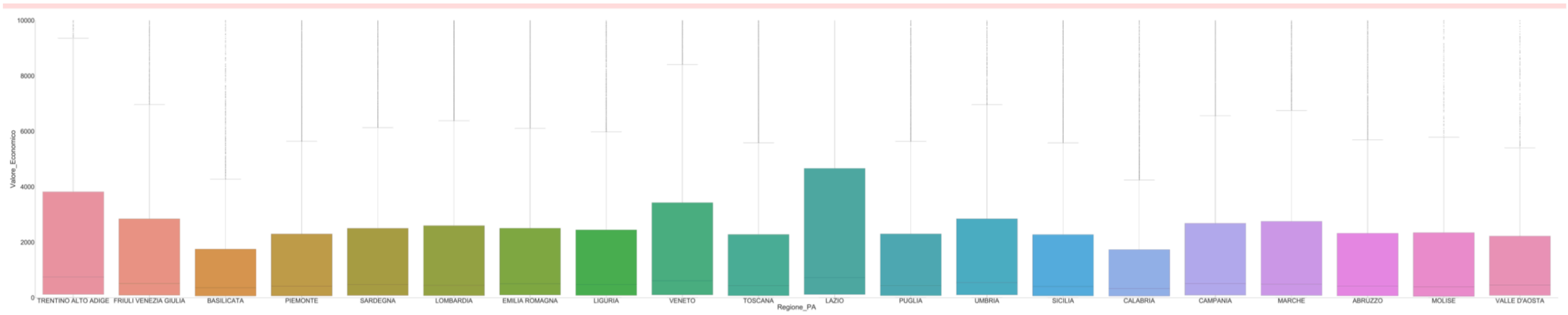
Lombardia e **Lazio** sono le regioni che hanno registrato il maggior numero di acquisti e/o negoziazioni dal 2018 ad oggi.



Acquisti & Negoziazioni (3)

Acquisti e negoziazioni effettuate dalle PA attraverso le varie piattaforme telematiche

Dal 2018 ad oggi: Lazio, Trentino Alto Adige e Veneto le regioni con maggiore spesa pubblica



Assunzione base

- Le PA non hanno al momento uno strumento che gli permetta di adottare un benchmark di riferimento nella valutazione del prezzo degli acquisti

Domanda di ricerca

È possibile prevedere il valore economico di un acquisto?

Obiettivo

- Creare un modello di previsione del valore economico di ogni acquisto.

Preparazione dataset

- Dopo aver eseguito le necessarie operazioni di pulizia dei dati, i file dei dati di acquisto e negoziazione sono stati combinati creando tra loro un dataset unico con **33 colonne** e **473 mila** righe.

```
df_finale.sample(2)
```

	Anno_Riferimento	Tipologia_Amministrazione	Regione_PA	Regione_Fornitore	Accordo_Quadro	Lotto	Bene_Servizio	Codice_CPV	Descrizione_C
43270	2019	AZIENDE ED ENTI TERRITORIALI DI SERVIZI PUBBLICI	TRENTINO ALTO ADIGE	LAZIO	non applicabile	FORNITURA DI LICENZE D USO IBM PASSPORT, DEL R...	SERVIZI DI MANUTENZIONE SOFTWARE	72267100-0	MANUTENZIC DI SOFTWARE TECNOLO DELL'I
3916	2020	AZIENDE ED ENTI DI SERVIZI SOCIO ASSISTENZIALI	EMILIA ROMAGNA	EMILIA ROMAGNA	non applicabile	non applicabile	CARTUCCE E TONER INK-LASER ORIGINALI	30125100-2	CARTUCCE TON

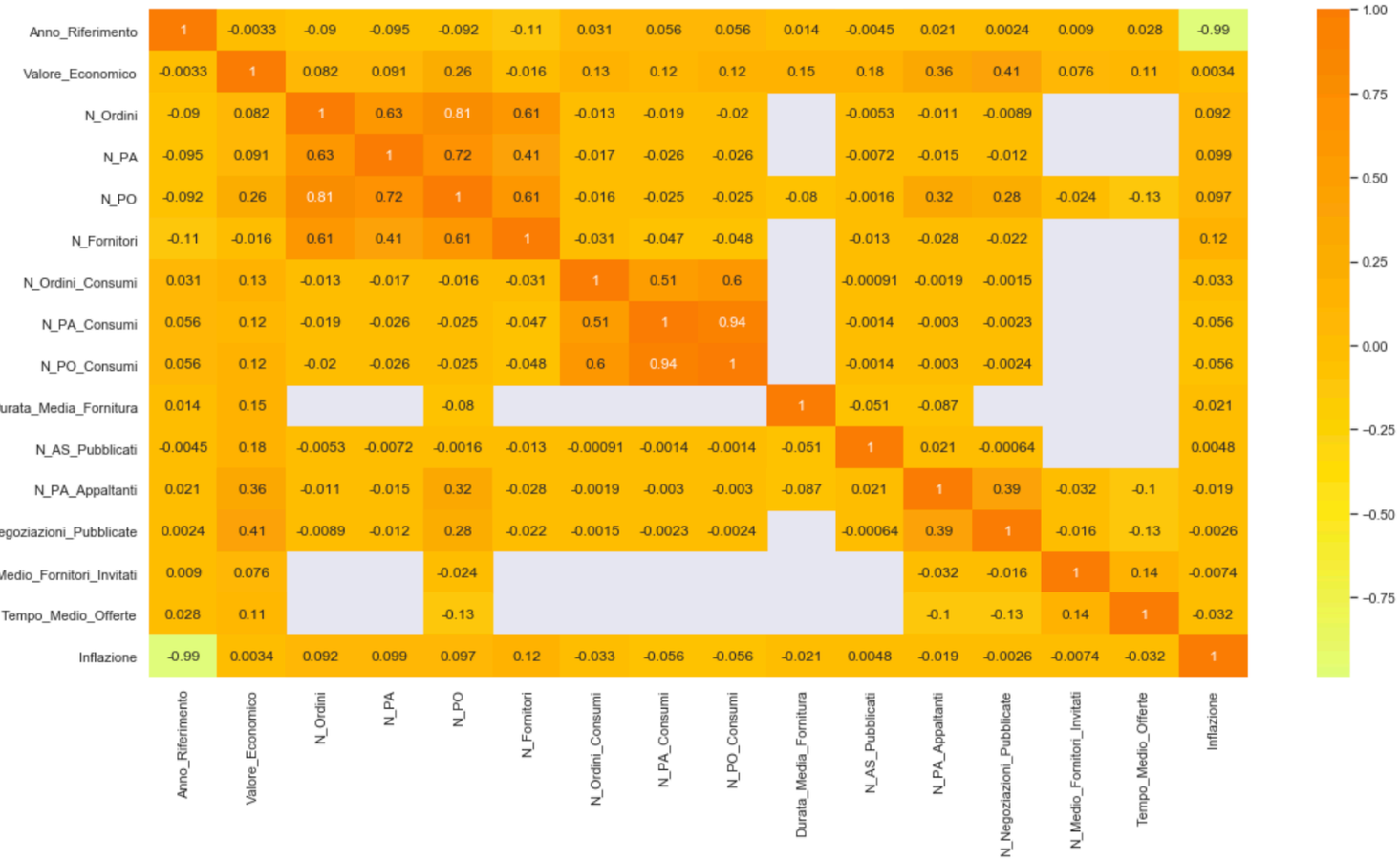
Analisi Esplorativa (1)

- Il dataset creato è rappresentativo di tutti e quattro gli strumenti di acquisto.
- AQ e SDA sono categorie sotto rappresentate.

Strumento_Acquisto	
AQ	2512
Convenzioni	65995
MEPA	401165
SDA	302

Analisi Esplorativa (2)

- E' stata effettuata un'analisi di correlazione tra le variabili numeriche, dando particolare rilievo alla variabile target: *Valore_Economico*

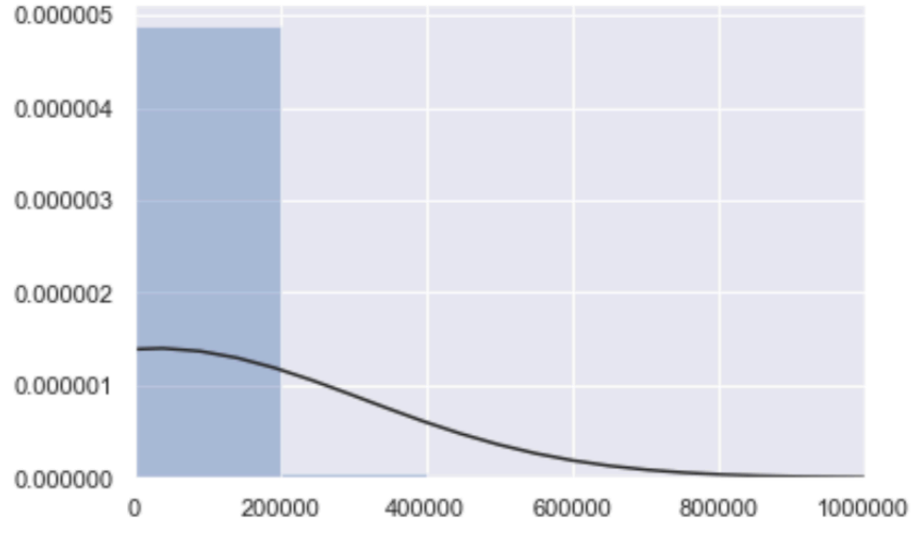


Analisi Esplorativa (3)

- La media della distribuzione è di ~€29,000, presentando una deviazione standard relativamente alta >€284,000 per il valore in oggetto. Questo a causa di alcuni acquisti con valori relativamente alti rispetto alla media (>€1Miliardo), che fanno scostare la curva verso destra.

- I valori fuori rango rappresentano acquisti effettuati tramite SDA.

	Valore_Economico
count	469974.00
mean	29022.97
std	284961.44
min	0.00
25%	81.15
50%	480.00
75%	2645.50
max	10034518.47



Train & Test

- Dopo aver convertito le variabili categoriche in numeriche ed aver standardizzato le stesse, il dataset finale risulta comprensivo di 2869 features

- Il dataset di riferimento è stato suddiviso per l'80% in training e per il 20% in test set.

Training del modello

- Per motivi di memoria e capacità computazionale si è estratto un campione random comprensivo del 10% del dataset in oggetto.

- Il fit dei dati è stato eseguito comparando una regressione lineare ed un *random forest regressor*, con parametri di default.

Model Selection

- Analizzando le performance a livello di *root mean squared error* si evince come *il modello di random forest* si presta molto meglio per questa task rispetto alla regressione lineare.

Regressione Lineare

```
Mean Absolute Error: 231941269558784384.000000
Mean Squared Error: 14928419974937522399136464246881124352.000000
Root Mean Squared Error: 3863731353877689344.000000
```

Random Forest

```
Mean Absolute Error: 32807.336143
Mean Squared Error: 43677962562.051163
Root Mean Squared Error: 208992.733276
```

Random Forest

- L'analisi del R^2 mostra come il primo modello creato riesce ad ottenere il fit del ~35% dei dati sul test set e >60% sul training set, mostrando dei possibili segni di overfitting.

- In questo caso l'overfitting potrebbe essere ottimizzato utilizzando più dati e/o regolarizzando i parametri utilizzati.

- Date le limitate capacita computazionali a disposizione, è stato utilizzato il sistema di *cross validation* per la ricerca dei parametri ideali.

```
Parameters currently in use:

{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'mse',
 'max_depth': 5,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': -1,
 'oob_score': False,
 'random_state': 1,
 'verbose': 0,
 'warm_start': False}
```

Regolarizzazione dei parametri del modello

- Il metodo di *cross validation* è stato applicato utilizzando l'applicativo GridSearchCV.

- Questo ci ha permesso di analizzare le performance di 98 modelli diversi di random forest

- Si è constatato un miglioramento di R^2 pari al 6%

```
train R squared score 0.854, forest R squared test score: 0.412
```

```
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.401, total= 3.6min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.367, total= 3.5min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.538, total= 3.2min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=1, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.395, total= 3.9min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.358, total= 3.0min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.428, total= 3.3min
[CV] bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False
[CV]  bootstrap=True, max_depth=13, min_samples_leaf=3, min_samples_split=2, n_estimators=20, oob_score=False, score=
0.418, total= 3.3min
```

Lezioni apprese ed opportunità

- La corretta preparazione del dataset rappresenta uno dei core dello studio in oggetto, richiede tempo e precisione di analisi.

- Una conoscenza più approfondita del settore delle PA e dei processi sottostanti, avrebbe sicuramente permesso di ottenere degli output più efficaci.

- Si potrebbero combinare altre sorgenti dati presenti in indice PA e ANAC.

- Support vector regressor e modelli di deep learning andrebbero applicati per compararne le performance rispetto a quelli utilizzati.

- Maggiori capacità computazionali consentirebbero di eseguire il training sull'intero dataset al posto di un sample.

- Source bias è un problema da approfondire. Infatti, per una accuratezza maggiore potrebbero essere creati 4 modelli diversi, uno per ogni strumento di acquisto (MePa, AQ, CON, SDA).

Considerazioni

- Il valore economico totale di ogni acquisto effettuato dalla PA dipende da numerose variabili tra cui il numero di ordini, il tipo di bene o servizio ordinato, la regione dove risiede la PA, ecc. Poter prevedere il valore economico di ogni acquisto potrebbe permettere alle varie amministrazioni di prevedere la spesa da affrontare in anticipo e quindi di poter pianificare e programmare il proprio budget in modo più efficiente.