Author: Fiorentino Nicola - 705055
e-mail: n.fiorentino3@studenti.uniba.it

Brazilian e-Commerce Platform: Optimize Order Fulfillment Times

I. Business Understanding

Il progetto realizzato prevede l'analisi di un dataset contenente le informazioni anonime di 100.000 ordini eseguiti fra il 2016 e il 2018 sulla piattaforma di e-commerce Olist Store, operante su territorio brasiliano. Sono forniti dati su stato degli ordini, prezzi, tempi di consegna, caratteristiche dei prodotti e recensioni degli acquirenti.

• *Business Objective*

L'obiettivo prefissato è l'ottimizzazione dei tempi di evasione degli ordini, al fin di ridurre l'attesa dell'acquirente e incrementarne la soddisfazione complessiva. Una previsione di quali ordini siano suscettibili di subire ritardi consentirebbe di intervenire sull'aspetto logistico, individuando una priorità nell'iter di approvazione o scegliendo un servizio di consegna rapida. Si desiderano, inoltre, scoprire eventuali fattori all'origine dei ritardi, in modo da adottare strategie di correzione adeguate.

• *Data Mining Goals*

Basandosi su tipologia e prezzo dei prodotti, costi di spedizione, data d'acquisto e luogo di residenza dell'acquirente, si apprenderà un modello di classificazione che consenta di individuare ordini con tempi di evasione superiori alla media (task predittivo). In secondo luogo si punterà ad individuare pattern significativi, che si manifestino in ordini caratterizzati da ritardi nella consegna (task descrittivo).

II. Data Understanding

• *Describe/Explore Data*

Il dataset considerato si compone di 25 attributi, suddivisi in categorici (A), numerici (#), date (📅) e chiavi identificative (🔑). I 100.000 record al suo interno fanno riferimento a 96.000 ordini distinti. Ogni ordine contiene uno o più articoli, eventualmente gestiti da venditori differenti, e risulta descritto da un numero di record proporzionale ai diversi prodotti di cui è composto. I prezzi sono espressi in real brasiliani (1 EUR = 6,75 BRL). Di seguito vengono riportate alcune statistiche associate alle diverse feature.

order_id

96264
unique values

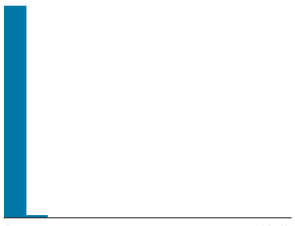
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	96.3k	
Most Common	5a3b1c29a...	0%

order_status

delivered	98%
shipped	1%
Other (1220)	1%

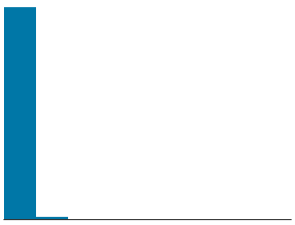
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	7	
Most Common	delivered	98%

order_products_value



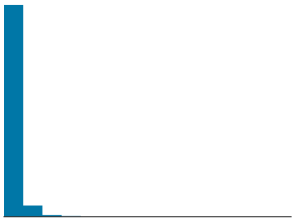
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	129	
Std. Deviation	195	
Quantiles	2	Min
	47	25%
	83.7	50%
	146	75%
	13.4k	Max

order_freight_value



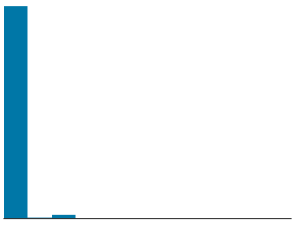
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	21.7	
Std. Deviation	20.1	
Quantiles	0	Min
	13.5	25%
	16.8	50%
	23	75%
	1.56k	Max

order_items_qty



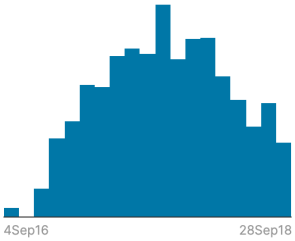
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.1	
Std. Deviation	0.46	
Quantiles	1	Min
	1	25%
	1	50%
	1	75%
	20	Max

order_sellers_qty



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.14	
Std. Deviation	0.71	
Quantiles	1	Min
	1	25%
	1	50%
	1	75%
	30	Max

order_purchase_timestamp

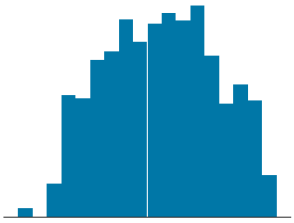


Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	4Sep16	
Mean	20Nov17	
Maximum	28Sep18	

order_aproved_at

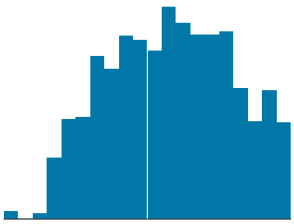
Valid	100.0k	100%
Mismatched	0	0%
Missing	18	0%
Minimum	15Sep16	
Mean	21Nov17	
Maximum	28Sep18	

order_estimated_delivery_date



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	4Oct16	
Mean	15Dec17	
Maximum	13Dec18	

order_delivered_customer_date



Valid	97.6k	98%
Mismatched	0	0%
Missing	2405	2%
Minimum	11Oct16	
Mean	4Dec17	
Maximum	3Oct18	

customer_id

96264
unique values

Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	96.3k	
Most Common	be1c4e52b...	0%

customer_city

Sao Paulo	14%
RIO DE JANEIRO	7%
Other (78864)	79%

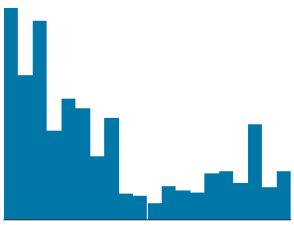
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	4121	
Most Common	Sao Paulo	14%

customer_state

SP	41%
RJ	13%
Other (45494)	45%

Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	27	
Most Common	SP	41%

customer_zip_code_prefix



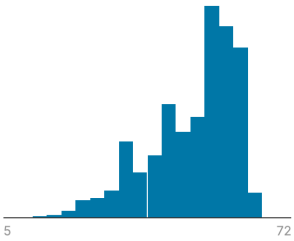
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	354	
Std. Deviation	298	
Quantiles	10	Min
	116	25%
	250	50%
	590	75%
	999	Max

product_category_name

cama_mesa_banho	10%
beleza_saude	8%
Other (81626)	82%

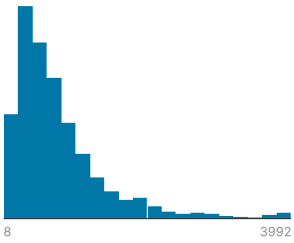
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	71	
Most Common	cama_mesa...	10%

product_name_lenght



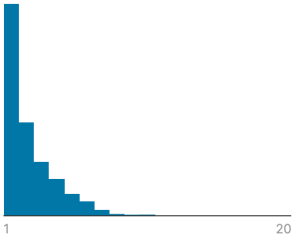
Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	48.8	
Std. Deviation	10.1	
Quantiles	5	Min
	42	25%
	52	50%
	57	75%
	72	Max

product_description_lenght



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	779	
Std. Deviation	665	
Quantiles	8	Min
	340	25%
	591	50%
	978	75%
	3992	Max

product_photos_qty



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.28	
Std. Deviation	1.75	
Quantiles	1	Min
	1	25%
	2	50%
	3	75%
	20	Max

product_id

24447
unique values

Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	24.4k	
Most Common	99a4788cb...	1%

review_id

96264
unique values

Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Unique	96.3k	
Most Common	5a3b1c29a...	0%

review_score

Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Mean	4.05	
Std. Deviation	1.36	
Quantiles	1	Min
	4	25%
	5	50%
	5	75%
	5	Max

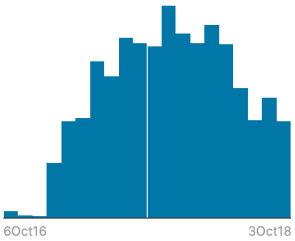
review_comment_title

[null]	92%	
Recomendo	0%	
Other (7968)	8%	
Valid	8262	8%
Mismatched	0	0%
Missing	91.7k	92%
Unique	3365	
Most Common	Recomendo	0%

review_comment_message

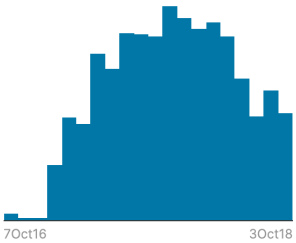
[null]	58%	
muito bom	1%	
Other (41917)	42%	
Valid	42.5k	42%
Mismatched	0	0%
Missing	57.5k	58%
Unique	33.8k	
Most Common	muito bom	1%

review_creation_date



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	6Oct16	
Mean	3Dec17	
Maximum	3Oct18	

review_answer_timestamp



Valid	100.0k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	7Oct16	
Mean	7Dec17	
Maximum	3Oct18	

- *Verify Data Quality*

- Le chiavi identificative *order_id*, *customer_id, review_id* appaiono ridondanti, associando un valore unico a ciascun ordine.
- I diversi attributi di tipo data presentano valori in formati differenti.
- Le date di approvazione o consegna di un ristretto numero di ordini non sono disponibili.
- Meno del 10% degli ordini possiede un commento rilasciato dall'utente sotto forma di titolo. Solo il 42% degli ordini è associato a una recensione dell'utente.
- Alcuni nomi di città o recensioni presentano al proprio interno apici o doppi apici che minano la struttura del dataset.
- I dati fanno riferimento ad ordini eseguiti fra il 2016 e il 2018.

III.Data Preparation

• *Select Data*

Il dataset è stato ottenuto campionando il database della piattaforma di e-commerce, pertanto non si reputano indispensabili ulteriori operazioni di sampling. Si prenderanno in considerazione i soli ordini per i quali sia disponibile la data di consegna al cliente finale (98% del totale). Inoltre, al fin di garantire una corrispondenza biunivoca tra ordini e record, verranno esaminati i soli ordini contenenti articoli di stessa natura (97% del totale). Informazioni relative ad eventuali recensioni saranno trascurate, poiché aggiunte in una data successiva all'immissione dell'ordine. Alla luce degli obiettivi prefissati, si assumeranno come rilevanti le seguenti feature:

- *order_products_value*
- *order_freight_value*
- *order_items_qty*
- *order_purchase_timestamp*
- *order_delivered_customer_date*
- *customer_city*
- *customer_state*
- *customer_zip_code_prefix*
- *product_category_name*
- *product_name_length*
- *product_description_length*
- *product_photos_qty*
- *product_id*

• *Clean Data*

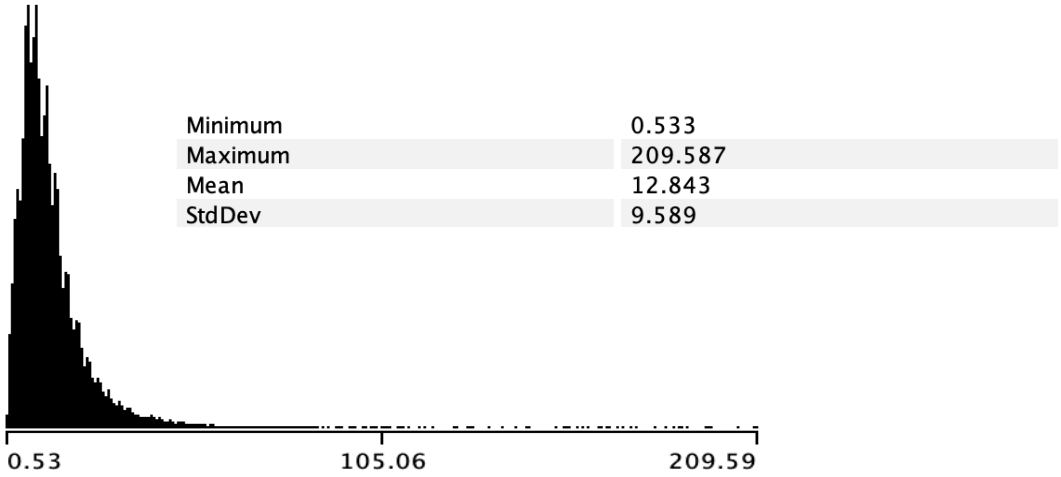
- Attraverso l'uso di un text editor sono stati rimossi apici e doppi apici da nomi di città e recensioni, al fin di consentire la corretta analisi del dataset.
- Mediante l'uso di espressioni regolari tutti i valori di tipo data sono stati ricondotti al formato "yyyy-MM-dd'T'HH:mm:ss".
- Sono stati rimossi alcuni record associati a date non riconosciute come valide.

• *Construct Data*

- L'attributo *order_delivered_customer_date* è stato trasformato in una target feature binaria *order_delay,* ad indicare un tempo di evasione dell'ordine superiore alla media:

  ```
  True <=> (order_delivered_customer_date − order_purchase_timestamp) > 13 days
  ```

  Di seguito vengono fornite alcune statistiche sui giorni intercorsi tra l'immissione di un ordine e la sua consegna al cliente.



- L'attributo *order_purchase_timestamp* è stato trasformato nella feature numerica *order_purchase_day,* ad indicare il giorno dell'anno in cui l'acquisto è stato effettuato.
- Tra i 4121 valori distinti dell'attributo *customer_city*, i meno frequenti sono stati fusi in una singola categoria "other". Una frequenza minima pari a 250 consente di ridurre i valori distinti a 46.
- Tra i 27 valori distinti dell'attributo *customer_state*, i meno frequenti sono stati fusi in una singola categoria "other". Una frequenza minima pari a 1000 consente di ridurre i valori distinti a 13.
- Tra i 71 valori distinti dell'attributo *product_category_name*, i meno frequenti sono stati fusi in una singola categoria "other". Una frequenza minima pari a 300 consente di ridurre i valori distinti a 30.
- L'attributo *product_id* è stato trasformato nella feature numerica *product_purchase_frequency,* ad indicare il numero di ordini in cui l'articolo è presente.

• *Integrate Data*

Le unità analizzate corrispondono al sottoinsieme dei 93.000 ordini descritti da singoli record del dataset. Per agevolare la comprensione si è realizzata un'integrazione con il dataset accessorio *product_category_name_translation.csv*, al fin di convertire i valori dell'attributo *product_category_name* nei rispettivi equivalenti in lingua inglese.
Il tentativo di integrazione con un secondo dataset contenente informazioni relative alla collocazione geografica dei venditori è stato abbandonato, non apportando alcun miglioramento all'accuratezza predittiva.

IV. Modeling

• *Select Modeling Technique*

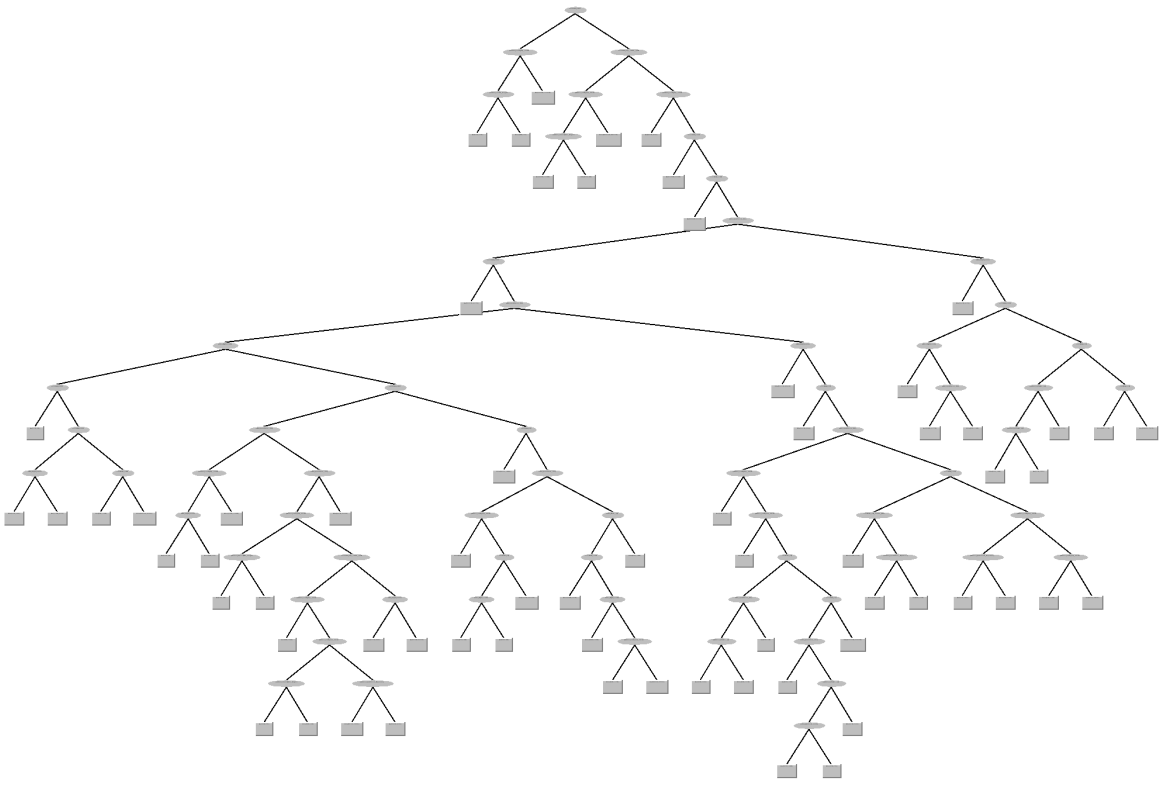
Il task di classificazione è stato eseguito ricorrendo all'uso di alberi decisionali. Tale modello è stato adottato in virtù del limitato costo computazionale richiesto in fase di apprendimento e classificazione e per la capacità di gestire attributi ridondanti o di

rilevanza limitata. Per non incorrere in fenomeni di overfitting si è fissato un limite minimo al numero di istanze ammissibili per nodo foglia, adottando tecniche di pruning che limitassero la complessità del modello. Approcci di tipo instance-based sono stati scartati a causa della gravosità in termini computazionali. Analogamente, sono stati esclusi classificatori di tipo Naïve-Bayes in virtù della dipendenza esistente tra i diversi attributi.

Per il task di pattern discovery si è fatto ricorso a regole di associazione la cui conseguenza fosse rappresentata da valori positivi della classe *order_delay*. Non essendo praticabile un approccio di tipo brute-force, l'estrazione delle regole è stata condotta con algoritmo Apriori, sfruttando la proprietà di anti-monotonia del supporto per la ricerca di itemset frequenti, usati come premessa di singole regole. La scelta dell'algoritmo è stata dettata dalla necessità di utilizzare class association rules e gestire attributi di tipo categorico.

• *Build Model*

L'apprendimento dell'albero decisionale è stato affidato all'algoritmo C4.5 messo a disposizione dal software di machine learning Weka. L'algoritmo sfrutta il concetto di information entropy, selezionando, in corrispondenza di ciascun nodo, l'attributo che garantisce massimo information gain. Valutazioni empiriche hanno suggerito l'utilizzo di uno splitting a due vie, fissando a 25 il numero minimo di istanze per nodo fogliare. La potatura dell'albero è avvenuta mediante tecnica di subtree raising, consistente nella rimozione di un nodo e ridistribuzione delle sue istanze. Di seguito viene riportata la struttura del classificatore, che raggiunge un'accuratezza del 75% sul training set.



Nell'esecuzione del task di pattern discovery, sono stati utilizzati i soli attributi categorici secondo i requisiti tecnici dell'algoritmo Apriori. Per consentire l'estrazione di un numero adeguato di class rules, si è individuato un supporto minimo di 200 ordini e un valore di confidence superiore al 75%. Si è scelto di selezionare il sottoinsieme di regole associato a ritardi nella consegna, in quanto giudicato maggiormente significativo. Infine, a causa della dipendenza tra gli attributi *customer_city* e *customer_state* sono state eliminate eventuali regole superflue, la cui premessa fosse una semplice estensione di altre regole con uguale livello di confidence. Di seguito vengono riportate le prime 10 regole generate.

```
1. customer_state=RJ product_category_name=office_furniture (274) ==>
      order_delay=true (242)  Conf:0,88

2. customer_state=other product_category_name=garden_tools (252) ==>
      order_delay=true (214)  Conf:0,85

3. customer_city=Sao Luis (303) ==>
      order_delay=true (254)  Conf:0,84

4. customer_city=other customer_state=other product_category_name=furniture_decor (295) ==>
      order_delay=true (245)  Conf:0,83

5. customer_city=other customer_state=CE (681) ==>
      order_delay=true (563)  Conf:0,83

6. customer_city=other customer_state=other product_category_name=bed_bath_table (270) ==>
      order_delay=true (222)  Conf:0,82

7. customer_city=other product_category_name=office_furniture (773) ==>
      order_delay=true (635)  Conf:0,82

8. customer_city=other customer_state=other product_category_name=computers_accessories (339) ==>
      order_delay=true (274)  Conf:0,81

9. customer_state=other product_category_name=bed_bath_table (318) ==>
      order_delay=true (257)  Conf:0,81

10. customer_state=other product_category_name=furniture_decor (373) ==>
      order_delay=true (300)  Conf:0,8
```

• *Generate Test Design*

Per la valutazione del modello di classificazione ci si è affidati a uno schema di k-fold cross validation, utilizzando un valore di k pari a 10. Sul dispositivo usato in fase di progettazione, il tempo impiegato per eseguire la valutazione è di circa 1 minuto. Sono stati esaminati: accuratezza del modello, MAE, RMSE, Precision, Recall, F-Measure.

• *Assess Model*

L'albero di decisione risulta costituito da 131 nodi, dei quali 66 fogliari. Di seguito vengono riportati i risultati della valutazione.

Correctly Classified Instances	74,6063%
Incorrectly Classified Instances	25,3937%
Mean absolute error	0,3625
Root mean squared error	0,427

Classified as –>	true	false
true	18455	15268
false	7760	49201

order_delay	TP Rate	FP Rate	Precision	Recall	F-Measure
true	0,547	0,136	0,704	0,547	0,616
false	0,864	0,453	0,763	0,864	0,810
Weighted Avg.	0,746	0,335	0,741	0,746	0,738

In relazione al task di pattern discovery, le prime 10 regole estratte presentano un valore di confidence superiore al 80%. Un incremento nel livello minimo di supporto consente di estrarre regole garantite da un numero maggiore di ordini, al prezzo di una riduzione del valore di confidence.

V. Evaluation

• *Evaluate Results*

Le prestazioni del modello individuato si mostrano significative, benché migliorabili: l'accuratezza del classificatore potrebbe essere raffinata integrando ulteriori informazioni, come la valutazione media di ciascun venditore. I pattern estratti illustrano come specifiche aree geografiche e categorie di prodotti siano inclini ad imbattersi in tempi di evasione degli ordini superiori alla media.