Università degli Studi di Cagliari
Facoltà di Scienze Economiche, Giuridiche e Politiche
Corso in Data Science, Business Analytics e Innovazione

LABORATORIO DI BIG DATA

REPORT PROGETTO FINALE:

"LINEAR REGRESSION OF FLIGHT PRICES"

Professore:
Prof. Giuliano Armano

A cura di:
Francesco Mussetti

Anno Accademico 21/22

INTRODUZIONE

Il dataset oggetto d'analisi è tratto dal sito Kaggle.com e contiene informazioni relative alle caratteristiche dei Voli Aerei effettuati tra le principali città dell'India tra l'11 Febbraio al 31 Marzo 2022.
L'obiettivo di tale progetto è quello dell'applicazione dell'algoritmo statistico di "Regressione Lineare" al fine di addestrare il set di dati e prevedere la variabile target continua "Price" in funzione delle restanti Variabili Indipendenti.

IMPORT LIBRERIE, AVVIO SESSIONE SPARK E CARICAMENTO DEL DATASET

Come prima operazione è stato eseguito l'import delle varie librerie utili all'esecuzione delle funzioni applicate nel corso della stesura del codice. Tale area è stata costantemente aggiornata man mano che si rendevano necessarie le differenti librerie a seconda dell'obiettivo perseguito.
Principalmente ci si è avvalsi di librerie importate da "py.spark.sql" e da "py.spark.ml", oltre che al supporto fornito dalle librerie di "pandas", "seaborn" e "matplotlib" nella parte di Data Visualization.
A seguire è stata inizializzata la Sessione di Spark e tramite il suo reader è stato eseguito il caricamento del file CSV "Dataset_Voli", reso disponibile in formato DataFrame.

OVERVIEW DEL DATASET

A seguito del caricamento del Dataset, si è proceduto a una sua prima generale ispezione al fine di comprenderne le caratteristiche generali.
Con l'applicazione di semplici comandi si può notare come il Dataset presenti 300.153 osservazioni e 12 variabili totali, delle quali viene esplicitato contenuto e tipologia. Nel dettaglio:

1)_c0: variabile quantitativa che rappresenta semplicemente la numerazione delle singole osservazioni.
2) Airline: il nome della compagnia aerea. È una variabile categorica che ha 6 diverse compagnie aeree.
3) Flight: presenta le informazioni relative al codice di volo dell'aereo. È una variabile categorica.
4) Source City: Città da cui decolla il volo. È una variabile categorica che ha 6 città uniche.
5) Departure_time: variabile categorica ottenuta raggruppando i periodi di tempo in classi.
Fornisce informazioni sull'orario di partenza e dispone di 6 fasce orarie univoche.
6) Stops: variabile categorica con 3 valori distinti che memorizza il numero di fermate tra la città di origine e quella di destinazione.
7) Arrival_time: variabile categorica creata raggruppando gli intervalli di tempo in classi.
Ha sei fasce orarie distinte e fornisce le informazioni sull'orario di arrivo.
8) Destination_city: Città in cui atterrerà il volo. È una variabile categorica che ha 6 città uniche.
9) Class: una variabile categorica che contiene informazioni sulla classe del posto a sedere; ha due valori distinti: Business ed Economy.
10) Duration: una variabile quantitativa continua di tipo float che mostra la quantità di tempo complessiva necessaria per viaggiare tra le città in ore.
11)Days_left: questa è una variabile quantitativa espressa in giorni che viene calcolata sottraendo la data del viaggio dalla data di prenotazione.
12) Price: la variabile target fornisce le informazioni sul prezzo del biglietto in ruple Indiane.

DATA CLEANING, CASTING E RECONDING

A seguito di una prima panoramica generale del DataSet è stato possibile passare alla fase di ripulitura dello stesso, al fine di individuare ed eliminare valori che potessero compromettere la qualità delle analisi successive.

Si è dunque ricercata la presenza di valori Nulli e di Dati Duplicati che ha dato esito negativo in quanto il Dataset utilizzato era probabilmente già stato elaborato per permettere un suo utilizzo ottimale.

In questa fase si è proceduto anche all'eliminazione delle colonne "_c0" e "flight" in quanto entrambe non fornivano alcuna informazione aggiuntiva utile al fine della nostra analisi e creavano solo rumore di fondo.

Successivamente si è proceduto all' arrotondamento dei valori della variabile float "duration" e al suo casting ad integer al fine di avere dei valori più omogenei.

Stessa procedura è stata effettuata sui valori della variabile Target "Price", che è stata inoltre sottoposta ad un operazione di Recoding al fine di esprimere il Prezzo dei voli, prima espressi in ruple indiane e quindi aventi una grandissima variabilità, in multipli di 1000 e andando quindi a ricodificare la variabile target in "Price_in_K".

DATA VISUALIZATION

Una volta ripulito adeguatamente il Dataset, si è passati a una descrizione delle caratteristiche dello stesso attraverso un Analisi grafica delle variabili che ne permettesse una comprensione più facile e immediata.

A tal fine sono stati predisposti una serie di Grafici a Barre utili alla descrizione delle Variabili Qualitative, alcuni Istogrammi e Boxplot per le Variabili Quantitative e una serie di Boxplot che evidenziano la distribuzione delle Variabili Target rispetto alle variabili Indipendenti.

Con riguardo ai Grafici a Barre possiamo apprezzare immediatamente la frequenza delle varie classi delle variabili ed il confronto tra le stesse espresso in scala ordinale decrescente e ottenere interessanti informazioni:

- "Vistara" e a seguire "Air_India" sono senza dubbio le compagnie aeree che hanno effettuato un maggior numero di voli, mentre "Spice_jet" a confronto ne possiede un numero irrisorio.
- Gli aeroporti di "Mumbai" e "Delhi" sono quelli dove si decolla e si atterra più frequentemente, mentre "Chennai" ha un traffico più ridotto.
-La maggior parte dei voli parte di mattina e arriva di sera o di notte, oppure parte la sera e arrivano il mattino inoltrato, mentre pochissimi partono o arrivano la notte tardi e pochi arrivano il mattino presto.
- La stragrande maggioranza dei voli effettua un solo scalo, il resto sono diretti e pochissimi ne effettuano due o più.
- I voli classe "Economy" sono praticamente il doppio rispetto ai voli classe "Business"

Con riguardo agli Istogrammi ed ai Boxplot possiamo apprezzare immediatamente la frequenza delle Variabili Quantitative e quindi studiarne la Distribuzione ottenendo interessanti informazioni:

- Per la variabile "Duration" possiamo osservare una distribuzione Asimmetrica a Destra e una Media intorno al valore 11 ed osservare che il valore con più frequenza è 2 ore, evidente è inoltre la presenza di numerosi Outliers, in tal senso si è quindi proceduto a un conteggio dei valori superiori a 40 ore e riscontrando una numerosità esigua si è proceduto all'eliminazione di tali valori dalla variabile ed a una successiva verifica tramite nuovi grafici.
- Per la variabile "Days_left" possiamo osservare una distribuzione leggermente Asimmetrica a Sinistra, assenza di Outliers, possiamo vedere come sia più raro che i voli vengano prenotati pochi giorni prima del volo, mentre già dal 5 giorno antecedente mantengono una numerosità regolare.
- Per la variabile Target "Price_in_K" possiamo osservare una distribuzione completamente Asimmetrica a Destra e presenza di Outliers, dunque cosi come effettuato in precedenza per la "duration" si è proceduto a un conteggio dei valori superiori a 100 K e riscontrando una numerosità esigua si è proceduto all'eliminazione di tali valori dalla variabile ed a una successiva verifica tramite nuovi grafici.

Con riguardo ai BoxPlot che mettono a confronto le Variabili Indipendenti rispetto alla Variabile "Price_in_K" possiamo osservare che:
- "Vistara" e "Air_India" sono mediamente più care rispetto alle altre compagnie.
- L'aeroporto di Partenza e di Arrivo non influisce significativamente sul cambio del Prezzo medio, solo Delhi sembra essere leggermente più economica.
- L'orario di partenza non influisce troppo sul prezzo medio tranne che per i viaggi in partenza la notte tarda.
- Anche per l'orario di arrivo non si riscontrano grandi differenze, se non per gli arrivi a tarda notte o mattina presto.
-I voli con uno scalo sembrano essere più costosi, seguiti da quelli con due o piu scali, mentre quelli diretti risultano in media più economici.
- Evidente è la differenza di prezzo medio tra le classi "Economy" e "Business, il prezzo medio della "Economy" si aggira attorno alle 7000 ruple contro le 57000 della "Business".
- Rispetto alla durata del volo possiamo osservare che il prezzo medio è crescente all'aumentare delle ore di volo e si stabilizza per i voli di durata superiore alle 10 ore.
- Rispetto ai giorni di anticipo tra prenotazione e volo, possiamo osservare che il prezzo ha un andamento decrescente all'aumentare dei giorni di anticipo, nel dettaglio sono più costosi i voli prenotati con un anticipo di 1,2,3 giorni , poi decresce gradualmente fino al 17 giorno a seguito del quale il prezzo si stabilizza .

DATA TRASFORMATION

In seguito all'Analisi Grafica, si è proceduto alla trasformazione dei dati eseguendo nello specifico lo Scaling delle Variabili Quantitative e l'Encoding delle Variabili Categoriche.
Con riguardo allo Scaling della variabile "Duration", data la sua distribuzione Asimmetrica a Destra e presenza di Outliers si è optato per l'applicazione di un algoritmo di RobustScaler che risulta più adatto a tali caratteristiche.
Con riguardo allo Scaling della variabile "Days_left", data la sua distribuzione Non Normale ma con assenza di Outliers, si è optato per l'applicazione di un algoritmo di MinMaxScaler che risulta più adatto a tali caratteristiche.
Con riguardo all'Encoding delle Variabili Categoriche, essendo tutte Nominali si è optato per l'applicazione degli algoritmi di String Indexing e di One Hot Encoding che risultano più adatti a tali caratteristiche.

Al termine di ciascuna delle precedenti operazioni, si è proceduto alla eliminazione delle variabili originarie in modo da mantenere nel Dataset esclusivamente i valori trasformati utili all'applicazione degli algoritmi successivi.

ASSEMBLY "FEATURES" E SUDDIVISIONE DEL DATASET IN TRAIN E TEST SET

Una volta ottenuto il Dataset con i valori trasformati, si è proceduto ad un'operazione di Assembly delle Variabili Indipendenti che sono state accorpate in un'unica variabile rinominata "features" utile ad essere data come input nel successivo Modello di Regressione Lineare, ottenendo di fatto un nuovo Dataset composto esclusivamente dalla variabile target "Price_in_K" e dalla variabile "features".
Come da prassi si è poi proceduto alla suddivisione di tale Dataset in Train Set per un 70% ed in Test Set per il restante 30%.

APPLICAZIONE DEL MODELLO "PIENO" → (Y= "Price_in_K" ; X= All)

Si è infine proceduto con l'effettiva applicazione del Modello di Regressione Lineare utile alla predizione della variabile target "Price_in_K" utilizzando come preditore la variabile "features" contenente quindi tutte le Variabili Indipendenti.

I risultati ottenuti presentano ottimi valori di R^2 sia per il Train Set che per il Test Set, ovvero:

- R^2 per il Train Set = 0.9126143199834577
- R^2 per il Test Set = 0.9125960776751125

E' stato inoltre calcolato anche l'R^2 aggiustato per il Test Set, che è leggermente più basso ma conferma la bontà del modello.

- R^2adj per il Test Set = 0.9125669776310567

APPLICAZIONE DEL MODELLO → (Y= "Price_in_K" ; X= Airline,Class,Duration,Days_left)

Sulla base delle informazioni ottenute durante l'analisi del Dataset, si è voluto indagare sull'influenza maggiore che alcune Variabili Indipendenti hanno mostrato di avere sulla variabile Target rispetto alle altre. A tale scopo è stato implementato un secondo modello di Regressione Lineare utilizzando come preditore la variabile "features" contenente solo le variabili Indipendenti "Airline","Class","Duration"e "Days_left".

I risultati ottenuti presentano anche in questo caso ottimi valori di R^2 sia per il Train Set che per il Test Set, anche se leggermente inferiori al Modello "pieno", ovvero:

- R^2 per il Train Set = 0.8991198586717164
- R^2 per il Test Set = 0.8987408278051601

E' stato inoltre calcolato anche l'R^2 aggiustato per il Test Set, che è leggermente più basso ma conferma la bontà del modello.

- R^2adj per il Test Set = 0.8987318398725572

APPLICAZIONE DEL MODELLO → (Y= "Price_in_K" ; X= Class)

Sulla base dello stesso ragionamento, si è voluto implementare un ulteriore terzo modello di Regressione Lineare utilizzando come preditore unicamente la variabile indipendente "Class" che aveva mostrato il maggior grado di influenza rispetto alla variabile Target.

I risultati ottenuti presentano anche in questo caso ottimi valori di R^2 sia per il Train Set che per il Test Set, anche se inferiori al Modello "pieno" e minimamente inferiori al Secondo Modello, ovvero:

- R^2 per il Train Set = 0.8809854952106777
- R^2 per il Test Set = 0.8812127254747689

E' stato inoltre calcolato anche l'R^2 aggiustato per il Test Set, che è leggermente più basso ma conferma la bontà del modello.

- R^2adj per il Test Set = 0.8812114076076067

CONCLUSIONI

Dati i risultati ottenuti possiamo infine affermare che i modelli di Regressione Implementati presentano un elevato adattamento ai dati e riescono dunque a spiegare in maniera molto buona il fenomeno in oggetto. In altre parole, le variabili indipendenti utilizzate nei modelli riescono predire molto bene la variabile dipendente "Price_in_K".

Il Modello Pieno in tal senso presenta il miglior livello di R^2 aggiustato, pari a 0,912.. , ed è dunque da considerarsi il migliore in termini predittivi.

Vi è da evidenziare però che, come dimostravano già le caratteristiche di alcune variabili indipendenti in fase di Data Visualization, e come confermato successivamente dal Secondo Modello , gran parte della variabile "Price_in_K" è spiegata dalle variabili "Airline","Class","Duration" e "Days_left" che usate come preditori presentano un R^2 aggiustato di 0,898.. .

In tal senso è opportuno evidenziare come il Terzo Modello, con la sola variabile "Class" come preditore, riesca a raggiungere un R^2 aggiustato di ben 0,881… .