11/09/2021

Danielle sherman

CTO, Blackwell Electronics

Overview:

As per your request, I have collected key take-aways and lessons learned about data analytics. The following Python tools and functions were utilized during Exploratory Data Analysis: Pandas to clean and compute data, Numpy to split, segment, and compute data, Matplotlib to visualize data and create reports, and scikit-learn to build classification and prediction models.

Investigate Customer Buying Patterns:

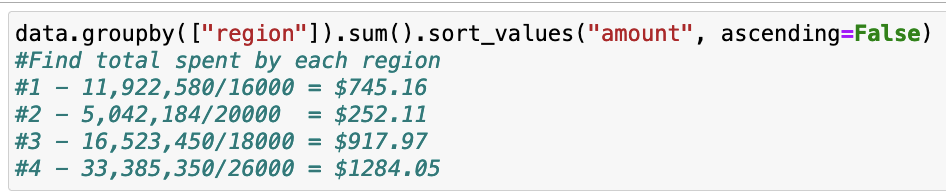
This task entailed answering the following questions:

**1.“Do customers in different regions spend more per transaction? Which regions spend the most/least?**

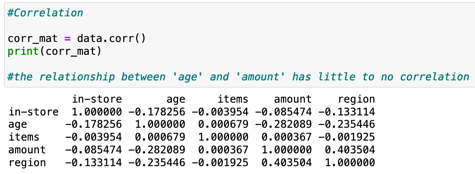
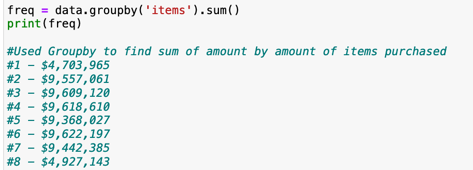
**2.“Is there a relationship between number of items purchased and amount spent?”**

This lesson was useful because it highlights the relationship between customer features. Using these relationships, the Marketing team can allocate funding with a strong degree of confidence.

1. I started by cleaning data using Pandas drop\_duplicates() function. I also checked for missing values with the isnull().sum function. It is useful to know the data types you are working with and make sure they are numeric with the dtypes function. This indicated that all features were int64, except amount, which was float64. To answer the question, I used groupby([“region”]).sum().sort\_values(“amount”) to find the total spent by each region. I further used groupby(“region”).count() to find the number of transaction per region. Next, I then took those values and divided each sum by the amount of transactions in each region.



1. I used groupby(‘items’).sum() to see the total amount spent by number of products. I also used a scatter plot to visualize this result. To check the strength of the relationship, I made a correlation table which indicated that there is a weak correlation between number of items purchased and amount spent.



Investigate Customer Demographics:

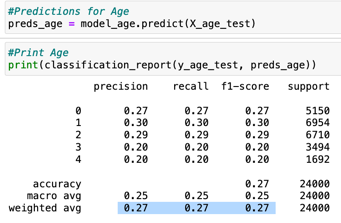
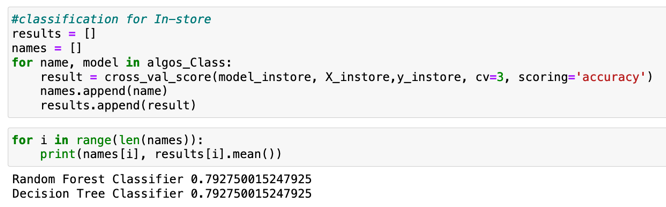
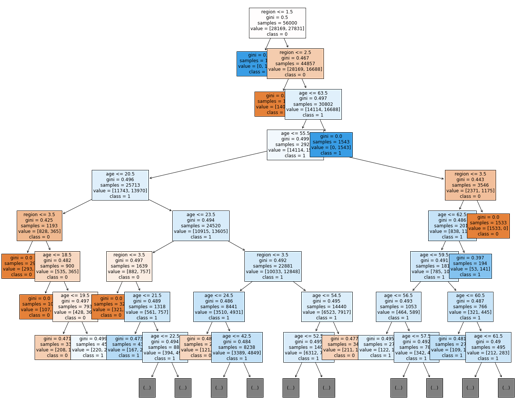
This task entailed answering the following questions:

**1. “Are there differences in the age of customers between regions? Can we predict the age of a customer in a region based on other demographic data?**

**2.“Is there any correlation between age of a customer and if the transaction was made online or in the store? Do any other factors predict if a customer will buy online or in our stores”**

This lesson was useful because it attempts to further flesh out customer demographics. Knowing your customer base and where they are located aids in marketing and helps grow the base.

1. I built Random Forest Classifier and Decision Tree Classifier models with sklearn in an attempt to predict outcomes. The most reliable model entailed using “Age” and “Region” as features with “In-store” as the dependent variable. This yielded an outcome of DTC=0.792750015247925, which is reliable. Both models aid in determining the strength of a prediction. However, the prediction score for “Age” was very low, hence it is more of a guess.



1. I utilized the same correlation table previously cited to answer this question. There is such a weak correlation between the age of a customer and if the transaction was made online or in the store.

Conclusion:

Task1 went smoother than Task2, this is down to the complexity of the tasks. Task1 was mainly simple math using Pandas and Numpy. Task2 was more complex in that it called for predictions using scikit-learn. In retrospect, I would not bin items that have a small number of variables. However, it worked well with Age by decade and splitting up Amount spent.

There was a lot of time wasted during the EDA. Although, the EDA is very helpful as it provides the basis for analysis, it can bloat your tasks, which can be confusing. Focus on the questions to be answered should be paramount to a project such as this.

In order for Blackwell Electronics to glean more information from their dataset, I suggest building regression models using “Amount” as the dependent variable. It was the only model that was not possible to build during tasks1-3.

GRANT JOHNSON