**Problem:**

Blackwell Electronics’ CTO Danielle Sherman has asked dive deeper in to specific customer demographics so we can better understand to whom to market and why.  **Infer** any insights about customer purchasing behavior, with below questions.

* Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
* We need to investigate Martin’s hypothesis: 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?

**Overview of the Data:** A text file was provided with data collected from recent online and in-store sales with a total 5 columns with row count = 80,000 raw, after removing the duplicates row count= 79,979. No irregularities where found with datatype mismatches. No need to fill in missing values or any data **Anomalies** using the box plot. All datatypes were numeric thus no need for ‘**Label and Coding**’ technique.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data Type** | **Null/Not Null** | **Description** |
| instore | integer | non-null | 0 or 1 |
| age | integer | non-null | Age of customer |
| items | integer | non-null | Number of items purchased |
| amount | float | non-null | Dollar amount |
| region | integer | non-null | * 1. region values |
| AgeInt | Integer | non-null | 6 separate classications(bins)  WHEN Age<=18 THEN 1  WHEN Age>18 and age<30 THEN 2  WHEN age >=30 and age<45 THEN 3  WHEN age >=45 and age<51 THEN 4  WHEN age >=51 and age<65 THEN WHEN age>=65 THEN 6 |

**Question #1**

Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?

**Results**

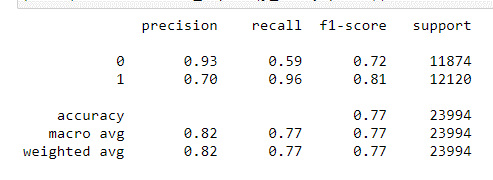
Age was a classification problem, after experimenting with different bins, I was NOT able to predict the age as it based on region or any other demographic data (online, amount, quantity, items). There is no predictive ability for age as it relates to this dataset. My accuracy was range from 40-50percent.

**Question #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?

**Results**

Yes, there is a relationship between age of customer and online. Other factors for example ‘items’ and ‘amount’ have no factors. Only age and region as features and ‘online’ as dependents. The precision is very high to make predications of **online/store purchases**, we can have issues with overfitting (not sure how to test). My predication, certain age groups (younger) shop online and certain regions are more prone to show online or at store.



**Separated variables into Features or X**

**Problem 1.** (In-Store, items, amount, and region) and dependents

**Problem 2**: ('age','region')

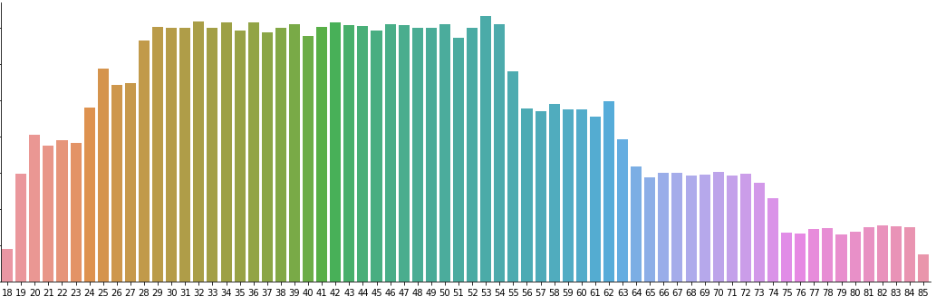
**Dependents or Y**

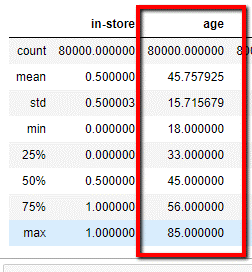
**Problem 1**: (Age)

**Problem 2**: Online or Store

Broke up the data into 2 folds: one for testing and one for training. First tested data for model selection and then real world. Measured error rate on a fold to pick the best model for **training**. Broke the age **into 6 separate classifications bins** (above).

**Age Metrics in dataset**





**Steps:**1. EDA (Load, clean and transform data)   
2. Visualize data and analyze  
3. Train, test split  
4. Build multiple models  
5 Score test data  
6. Evaluate model.

Cross Validation model then evaluate in real world. First part was used for model selection and 2nd fold was for the real world (real world data is never touched) testing. Use all folds, then

**Data🡪Train🡪Select Best🡪Test on data never touched🡪Estimate conservative.**

**Cross validation used for model selection, Used to select the model.**