|  |
| --- |
|  |
|  |
|  |
|  |

|  |
| --- |
| **Probabilistic Modeling project Documentation** |
| Alzahraa Mofed Mohamed 20191703001 Amira Barakat Ali 20191703002 Antony Nabil Naguib 20191703004 Aya Khaled Mohamed 20191703005 Nagwa Mohamed Anwar 20191703037   A |
| Project 2: Market Items Sales using Regression |

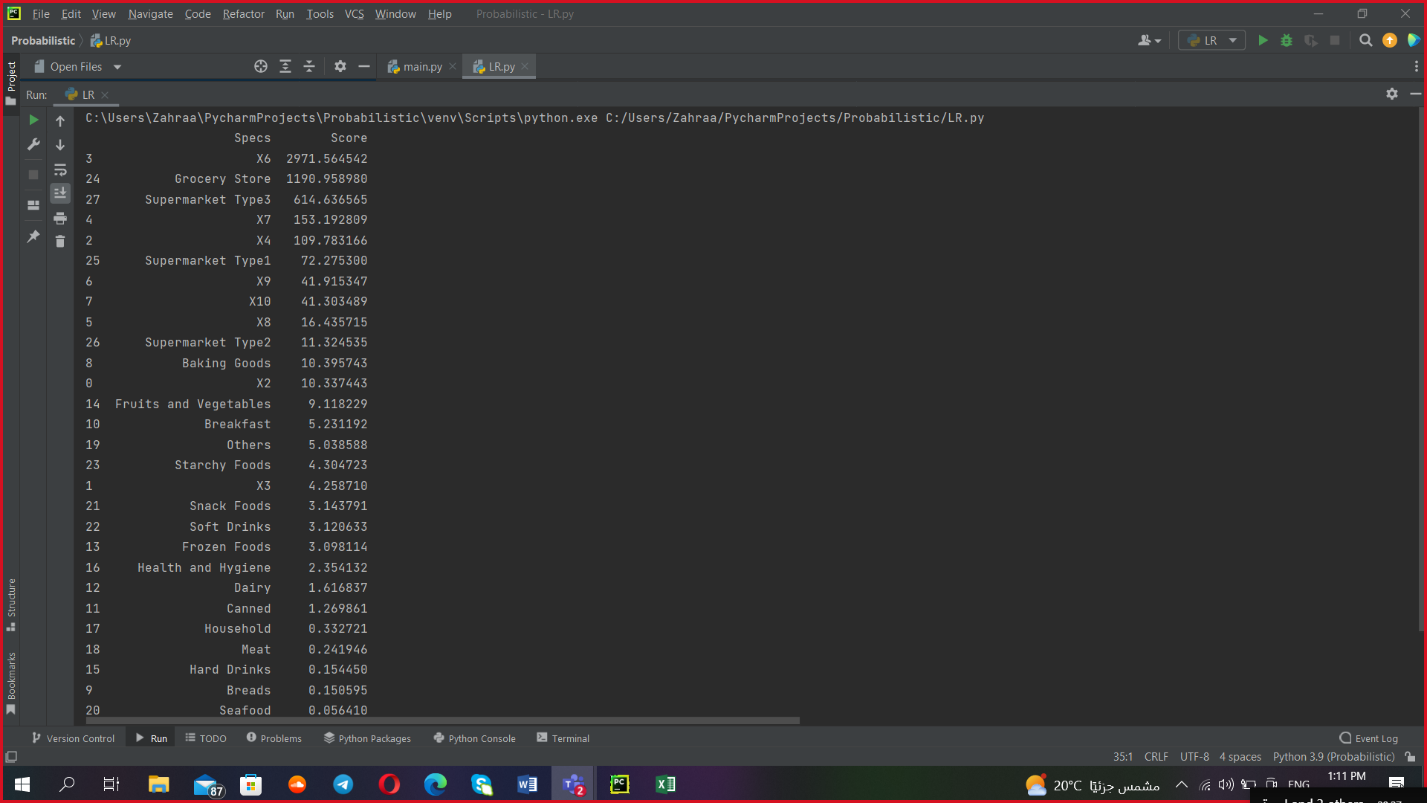
The Current Model:

## Data Preparation:

1. We started be missing null values using mode value.
2. We then dropped the duplicate rows.
3. We dropped the column X1 because it contains the ID’s.
4. In column X3, we replaced 'low fat', 'LF' with ‘Low Fat’ and we also replaced 'reg' with 'Regular'.
5. We performed label enconding on the columns X3, X7, X9 and X10 as they contain ordinal values.
6. We performed one-hot encoding using get dummies on X5 and X11.
7. We scaled the data using min-max scalar as the data doesn’t follow the normal distribution.

## Feature Selection:

1. We used The SelectKBest method to select the features according to the k highest score.



1. Based on the scores, we decided to select the features X6, X7, X4, X8, X9, X10, grocery store, Supermarket type1, Supermarket type2 and Supermarket type2 as they had the higgest scores.

## Used models:

We used the following models arranged from best to worst:

1. SVR- 1315.5537
2. Elasticnet- 1700.02815
3. Linear regression- 1701.51671
4. Catboost- 1703.02947
5. Ridge- 1703.43907
6. Lasso- 1704.19655
7. Decision trees- 1868.84852

Failed Trials:

## Removing Outliers:

1. Z-score: If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.
2. IQR:

* Find the first quartile, Q1.
* Find the third quartile, Q3.
* Calculate the IQR. IQR = Q3-Q1.
* Define the normal data range with lower limit as Q1–1.5\*IQR and upper limit as Q3+1.5\*IQR.

Both made the score worst so we decided against using them.

## Feature Scalling:

1. Robust scalar: removes the median and scales the data according to the quantile range.
2. MaxAbsScaler: Scales each feature by its maximum absolute value.
3. Standard scalar: removes the mean and scaling to unit variance.

Robust scalar and standard scalar performed badly because they need normally distributed data so, we ended up using min max scalar as it’s used for datasets not normally distributed such as the given dataset.

## Filling nulls using imputers:

The categorial features are only filled using the mode and encoded, then:

1. Filled the null values using imputer with estimator Bayesian Ridge
2. Filled the null values using imputer with estimator Catboost
3. Filled the null values using imputer with estimator Linear Regression