Using Customer Lifetime Value CLV to determine Superstore locations

Craig Sim

Agenda

- Task Outline
- Customer Lifetime Value
- Exploratory Data Analysis
- Model Creation
- Inference Results
- Further Considerations

Introduction

Task Outline

- From a selection Kaggle Datasets choose an interesting area, build a model and create a presentation

Approach

- I decided that a store or consumer dataset would be interesting, and it would also allow me to potentially create a Customer Lifetime Value CLV metric.
- Why CLV? CLV can has multiple useful applications for commercial and marketing purposes

Introduction: Why Customer Lifetime Value?

CLV is a crucial metric for businesses across various industries for several reasons:

Strategic Decision Making – customer acquisition value

Profit Maximisation – understanding the revenue a customer can generate

Customer Segmentation – helps segment customers on value

Forecasting and Planning – helps forecast revenue streams

Evaluation of Marketing Effectiveness – helps determine ROI

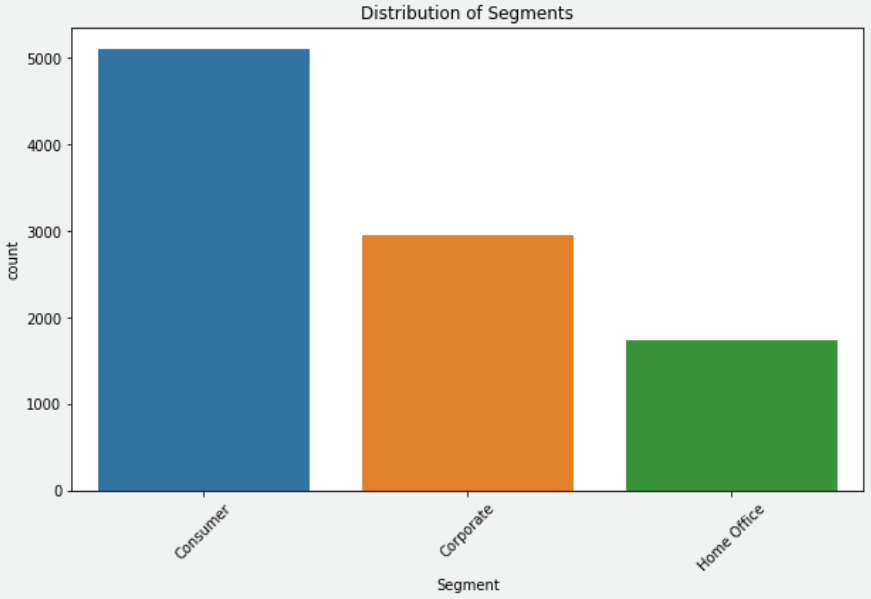
Can Give a Competitive Advantage – helps determine customer strategy

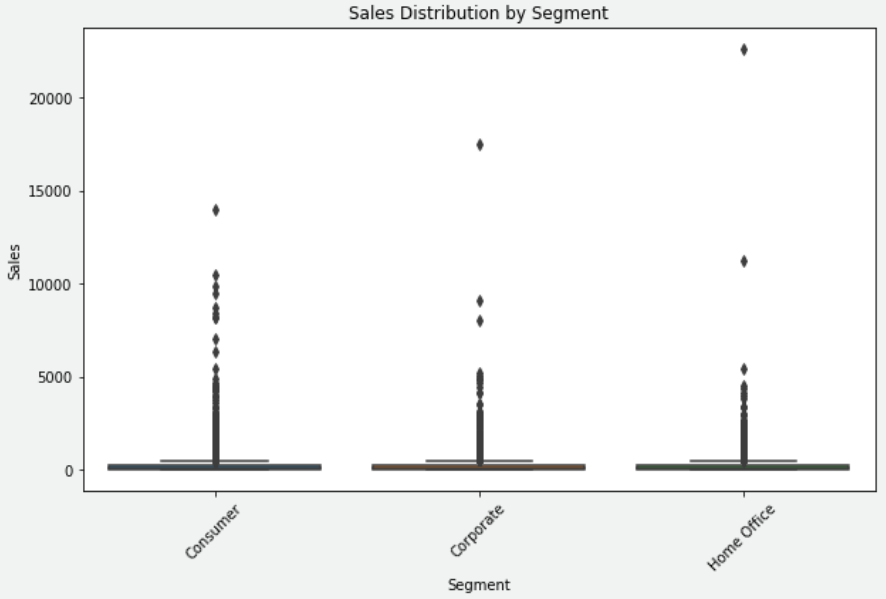
Superstore Dataset: Exploratory Analysis (1)

- The dataset contains a history of superstore orders within the USA
    - Data pertaining to Order and Shipping Dates, Customer ID and Name, Shipping Details, Location Details, Product Details and Sales revenue
    - There is also a feature that Segments the results by business area: Consumer, Corporate and Home Office
- There is sufficient data to derive a simple CLV metric using Sales, Order ID and Order Dates

Superstore Dataset: Exploratory Analysis (2)

- Key highlight from the dataset:
  - Order dates span a 3 year period from 2015 to 2018
  - Several customers have made multiple orders
  - California and New York are the top sales states, Wyoming and West Virginia the bottom
  - The West Region provides the most sales and the South region the least
  - Office Supplies are the top selling items
  - The Consumer segment is the top purchaser



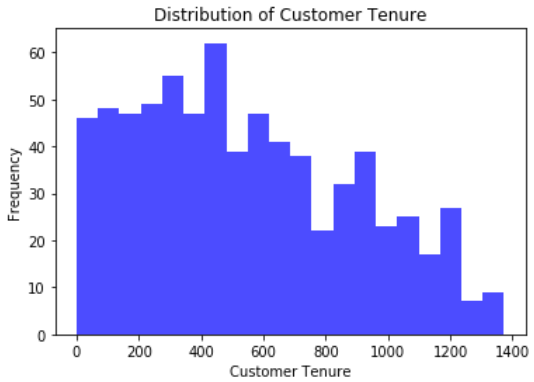


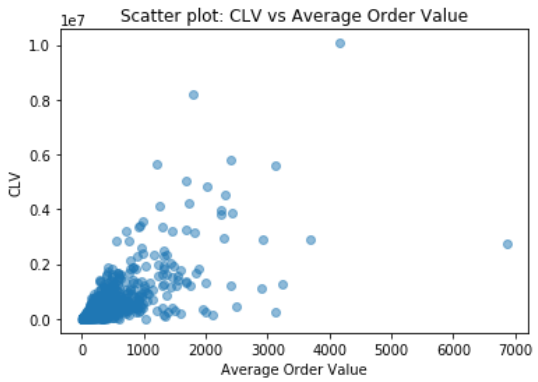
Create a CLV Model: Use it to predict best store locations

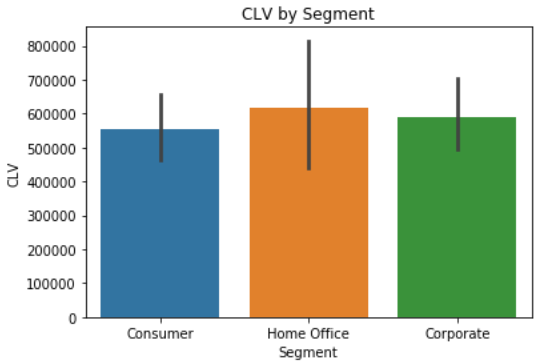
- We first need to calculate CLV from original superstore dataset:
  - CLV = average order value * frequency of purchases * customer tenure
  - For simplicity we have used average order value
  - Frequency of purchases is a count of individual customer purchases on different dates
  - Customer tenure is assumed to be length of time from first and last purchase

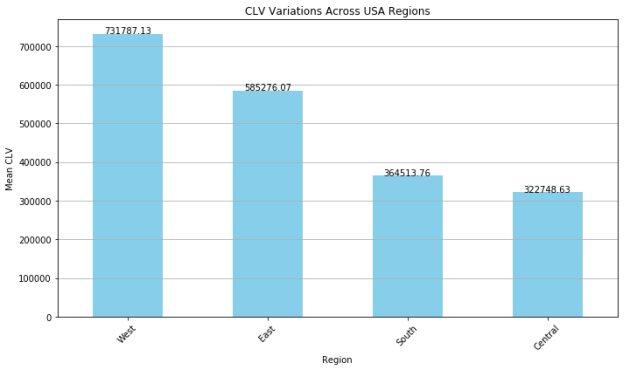
Using the original dataset calculate the above CLV metric and group the results by Customer_ID, Segment, Region and State

What does our CLV dataset tell us?

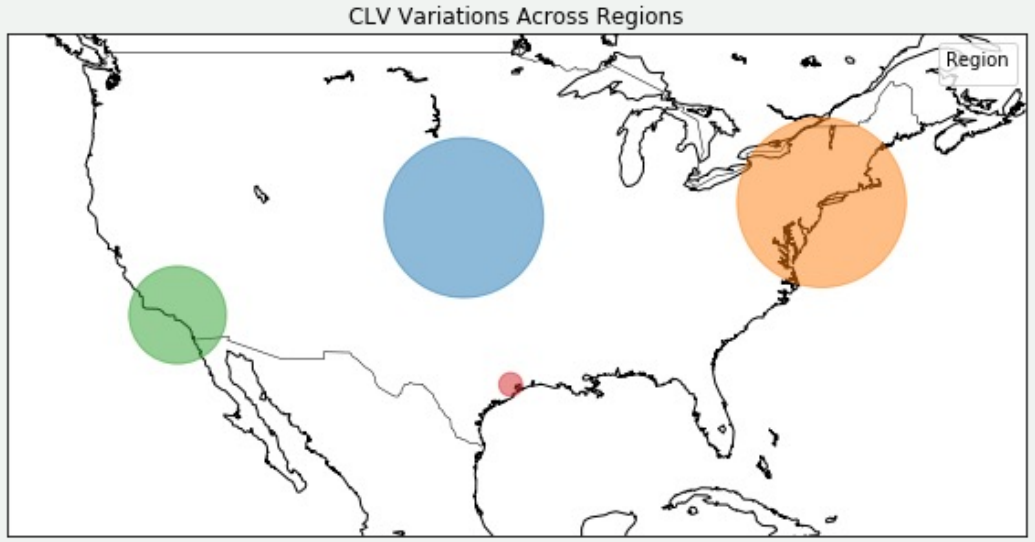


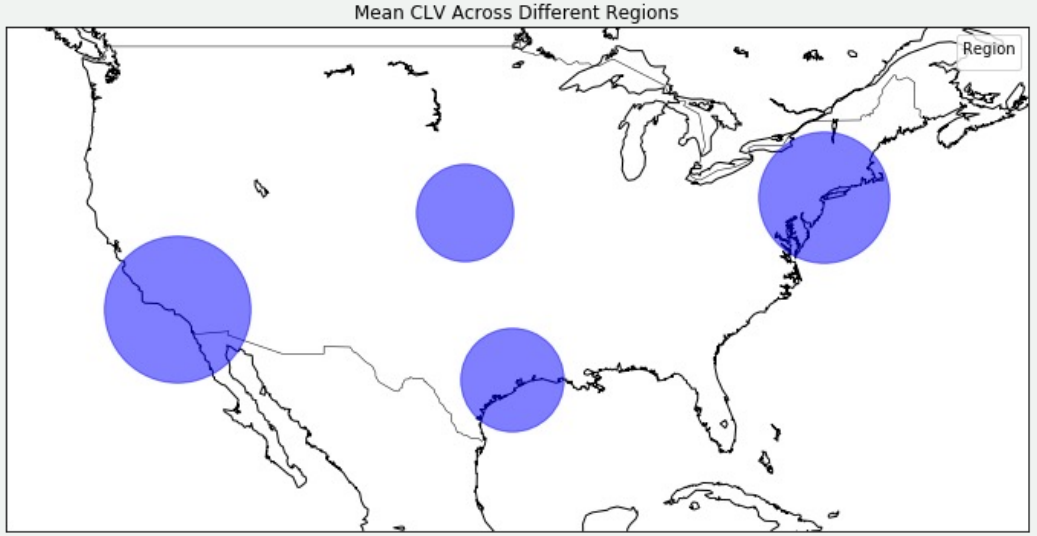






How does CLV vary geographically?





Model Choice

- As we are predicting a numerical value a regression model is appropriate
- After reviewing the dataset we:
    - Removed some extreme outliers
    - normalized the CLV target variable using a MinMaxScaler
    - Reduced features to segment and geographical features: Segment, Region and State, with CLV as target variable
- Two models were trained, XGBoost Regressor and a Linear Regressor
- The XGBoost model initially preformed best on the dataset, giving very low Mean Square Error on unseen data.

Inference Example

Business development want to assess the potential location of a new Superstore and have three US states in mind but only have budget for two stores. Where should they be located and what segment should they focus on?

If we run our model for the three locations and assess the CLV projected this will give Business Development some indication of where best to locate the stores and what to stock.

State	Segement	Average CLV Prediction
Washington	**Consumer**	$**958,008.12**
Washington	Corporate	$391,222.25
Washington	Home Office	$497,955.71
Pennsylvania	Consumer	$187,910.09
Pennsylvania	Corporate	$371,480.06
Pennsylvania	**Home Office**	$**689,144.68**
Ohio	Consumer	$170,754.84
Ohio	Corporate	$467,494.43
Ohio	Home Office	$**51180.13**

But... there's a problem

- We have a small dataset and this has potentially caused an overfitting issue
- Comparison with Linear Regression model, Cross Validation and Hyperparameter tuning tells us that our model may be overfitted to the small dataset. We are outperforming Cross Validation and we only have 703 rows in the dataset.
- One potential resolution is more data and richer data
- Interestingly we see a negative R^2 test result which suggest the data may have a non-linear relationship, or more likely, be of insufficient size. It important to re-evaluate the model, the data, and the context in which it was applied to understand why the R^2 value is negative and how to proceed.

How would we scale the CLV model?

We've created a specific model to predict CLV, by segment, for US States

However, if we have a large and more feature rich dataset, particularly around demographics of the customer, then we will be able to create more sophisticated predictions.

A more useful CLV model would taking into account time variances and churn hence we would consider using a hazard curve based model segmented by CLV itself. If the underlying dataset is broader then a variety of marketing interventions and/or business development questions can be addressed

The final model should also be developed into production pipelines and run using big data capable cloud based solutions such as AWS Sagemaker, Google or Azure.