Customer Segmentation & Prediction

Michael Hansen, Keith John, Ujjwal Aggarwal, Daniel Chindris, Anay Misra, Sruthi Vadakuppa, Anvesha Nain, Ishan Junnarkar, Nitya Jhaveri, Saumya Satishkumar Patel

Introduction

AgReliant is an agriculture company that primarily focuses on selling corn and soybean seeds throughout the Midwest.

- As a company, AgReliant seeks to be more efficient with their sales to customers
- Our purpose was to use data science solutions to help AgReliant gain insights into their customer base
- Given datasets with customer demographics and purchase history, our team leveraged machine learning techniques to segment customers, and predict future customers behavior

Stage One: Data Familiarization

With multiple datasets (customer demographics, purchase history, external market intelligence), we first focused on data exploration and cleaning.

- Leveraged Python to clean missing values in the data and combine all each dataset into one fully cleaned file

Stage Two: Model Development

Using the cleaned data from stage one, our team researched and implemented a wide variety of machine learning models to try and derive information about customer behavior.

- Used clustering algorithms to segment customers into different groups based on their demographics
- Explored tree-based learning algorithms such as Random Forests and gradient boosting to predict future customer behavior

Stage 3: Model Visualization

To increase the interpretability of our predictive machine learning models, we generated visualizations using Python libraries and Tableau dashboards

- Generated business insights with the usage of Python visualizations and Tableau dashboards with model results

Machine Learning Models

Customer Segmentation

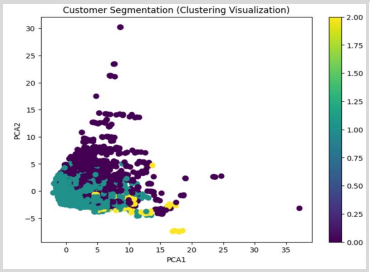
- BIRCH: a hierarchical clustering approach that incrementally builds a tree structure to summarize the dataset, enabling efficient clustering of large volumes of data while effectively managing noise and outliers.
- Silhouette Score: a metric used to evaluate the quality of clustering, measuring how well each data point fits within its assigned cluster compared to other clusters, with values ranging from -1 to 1 (We achieved a score of 0.60 - excellent clustering)

Customer Prediction

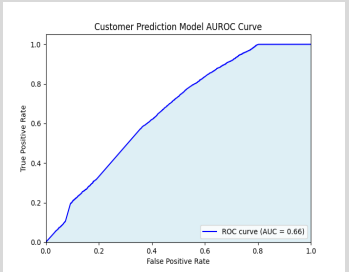
- RandomForest: Uses combinations of decision trees for classification. Averages results from multiple trees to reduce for overfitting and variance
- LightGBM: Builds on trees sequentially, with each tree building on the errors of the last one allowing for high-speed results.
- Feature Engineering: Created features taking aggregated findings from previous years and turning them into meaningful statistics to predict future years.

Results

The following image shows our clustering. It has a silhouette score of 0.5, which signifies that the clustering is reasonable.



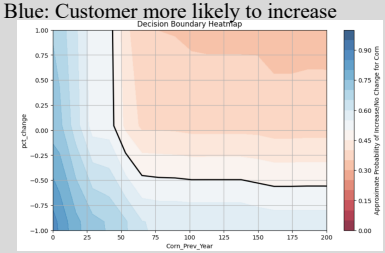
The AUROC curve illustrates how well the model distinguishes between buyers and non-buyers. It was modeled using Random Forest.



Results

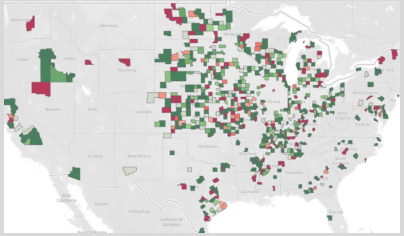
Predicted likelihoods that customers would increase their purchase volume from AgReliant in the year 2024 with 65% accuracy.

Red : Customer more likely to decrease
Blue: Customer more likely to increase



Created a map of how well the customer prediction model performed when predicting the 2024 sales year in Tableau

Green counties = model predicted customer behavior with high accuracy
Red counties = model predicted customer behavior with low accuracy



Conclusion / Future Goals

We successfully identified three distinct customer segments using the BIRCH clustering model. Additionally, our customer prediction model requires an accuracy assessment to evaluate its reliability and potential improvements. For sales forecasting, our Random Forest model achieved a 65% accuracy. In our future works we plan on enhancing this model through feature engineering, hyperparameter tuning and exploring more algorithms to improve our performance.

Acknowledgements: Patrick Sadtler, Jason Chia, The Data Mine Staff, AgReliant Staff supporting this project

References: sci-kit learn, numpy, pandas, python documentation