AGRO: An AGri-business Recommendation Optimization Engine for Sales Growth Decision-Support

Varsha Prabhakar, Rudraksh Syal, Yash Sharma, Zhi Dou, Matthew A. Lanham

Purdue University Krannert School of Management

prabhakv@purdue.edu; rsyal@purdue.edu; sharm364@purdue.edu; dou6@purdue.edu; lanhamm@purdue.edu

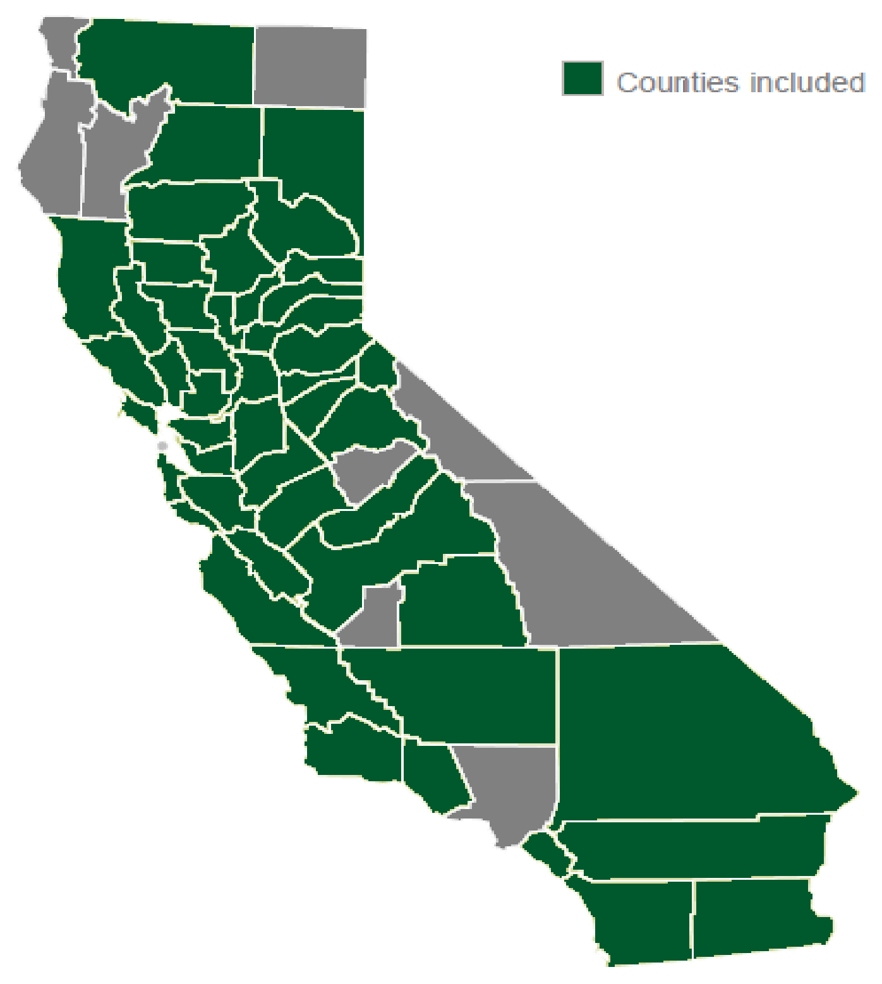
Abstract

In this study, we have built an intelligent agribusiness recommendation optimization engine – AGRO. It is developed to support an agri-science company's sales team to identify potential customers, understand market needs, and provide customized and accurate marketing recommendations regarding individual customers (farmers). Using a blend of multiple machine learning algorithms, this cross-validated engine ensures generated recommendations are optimized, inaccurate suggestions are minimized, and precisely predicts which segment of customers to focus with what category of products.

Introduction

The USA is one of the world's leading agricultural producers and supplier. California is the largest and the most dynamic agricultural marketplace in the US, and is responsible for over one third of the country's vegetables and over two third of the country's fruits and nuts. Growing more than 400 commodities of various crops all year round, approximately $47B dollars of total farm and ranch value is attributed to the state. For agribusiness companies, a deep understanding of farmers' need is critical for them to market their products and compete with other competitors.

Figure 1. Counties in consideration



In collaboration with an Agri-science company we use this public data with the firm's proprietary data to develop a cross-validated algorithm that provides strategic decision support to their sales team in providing assistance for whom to target and which assortment offering to provide them among an infeasible number of combinations. We built a Logistic Regression model to categorize the farmers into hot, warm, and cold leads. Further, we built another statistical model to dig deeper and predict the type of product a farmer is inclined to buy, out of Herbicide, Nematicide, etc.

Literature Review

We found that this topic has been attempted to be solved throughout the years and each article utilized a different algorithm to come up with their results. Most commonly, we found the use of Decision Trees, Bayes algorithm and K-means clustering. We used a simpler, more interpretable approach of Logistic Regression to create an optimized solution for our problem statement.

Study	Decision Tree	K-means	Generalized Linear Models	Bayesian	Support Vector Machine
K Liakos, P Busato, D Moshou, S Pearson, D Bochtis	✓	✓	✓	✓	
NG Hegde, S Mujumdar, SS Jambarmath, R Navada, Madhavi RP		✓		✓	
J. Ifft, R. Kuhns, K. Patrick.			✓	✓	✓
Our study	✓		✓		

Table 1. Literature review summary

Methodology

Data Set

The time-varying dataset of 12 million rows and 36 features involved the product usage for each permittee (farmer) along with the area of land treated by a product for a span of four years from 2014 to 2017. Further, the dataset also included additional features such as the type of product, county in use, type of the crop, application month and so on.

Exploratory Data Analysis & Pre-Processing

Due to the presence of just 3 numeric variables (out of 36) and quite a lot of junk variables, we had to put significant effort in curating the dataset. We visualized the total acre treated in each county to understand the market share enjoyed by our client and their competitors. We also did the quintessential steps of removing correlated features, eliminating linear combinations and non-zero variances.

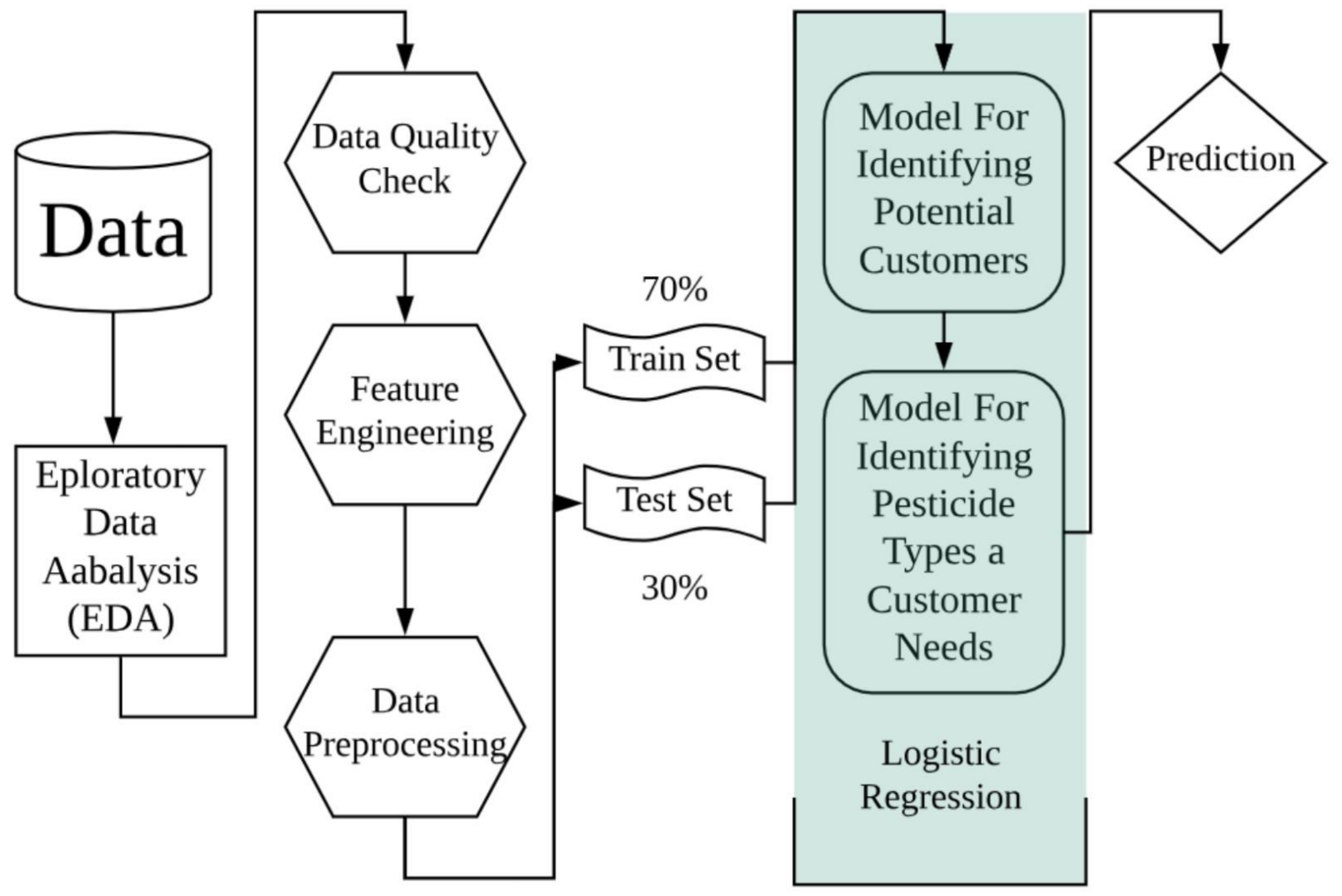

Data
Eploratory Data Aabalysis (EDA)
Data Quality Check
Feature Engineering
Data Preprocessing
70%
Train Set
Test Set
30%
Model For Identifying Potential Customers
Model For Identifying Pesticide Types a Customer Needs
Logistic Regression
Prediction

Figure 2. Schematic diagram of Model Methodology

Feature Engineering

Lack of significant number of features led us to prepare additional 24 features from the existing dataset such as brand loyalty, product usage, and couple of other weighted indexes over counties and farmers. This provided us with 10 significant features to efficiently predict our prospective leads in the industry.

Model Design

Owing to 32% of the uncaptured market share and identifying potential customers, we trained a Logistic Regression (Logit) model with a cut-off at 50% and 70:30 train-test data split. Further, we again used the same model to focus specifically on which pesticide type a customer needs and with what probability.

Methodology (Approach) Selection

We developed a couple of models such as Logit, Decision-trees, and Random Forest, but Logit was the best and the most interpretable approach of them all. There was no issue of overfitting and it gave us good results. For similar reason, we used the same approach while identifying pesticide type for a farmer.

Model Evaluation / Statistical & Business Performance Measures

We used Confusion Matrix approach to gauge Sensitivity (correctly predicting customers) and Specificity (correctly predicting non-customers). On this basis, we calculated AUC (Area under Curve) to strike balance between both these metrics.

Results

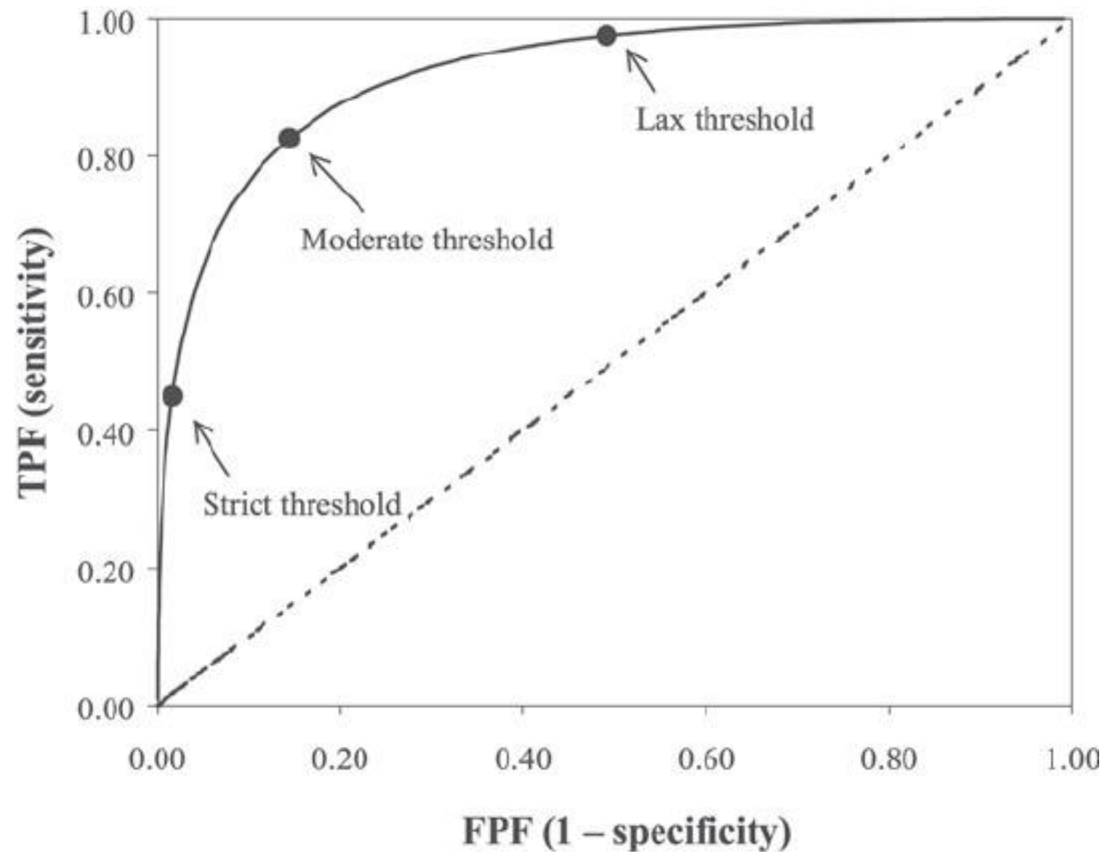


Figure 3. Model Accuracy to Predict Purchase

The first model which identifies probability of a farmer purchasing a company's product has a good accuracy and was not overfit. It can be used to identify the customers who did not make into three categories of lead hot, warm and cold. This would enable the sales team to find which new products are most desirable.

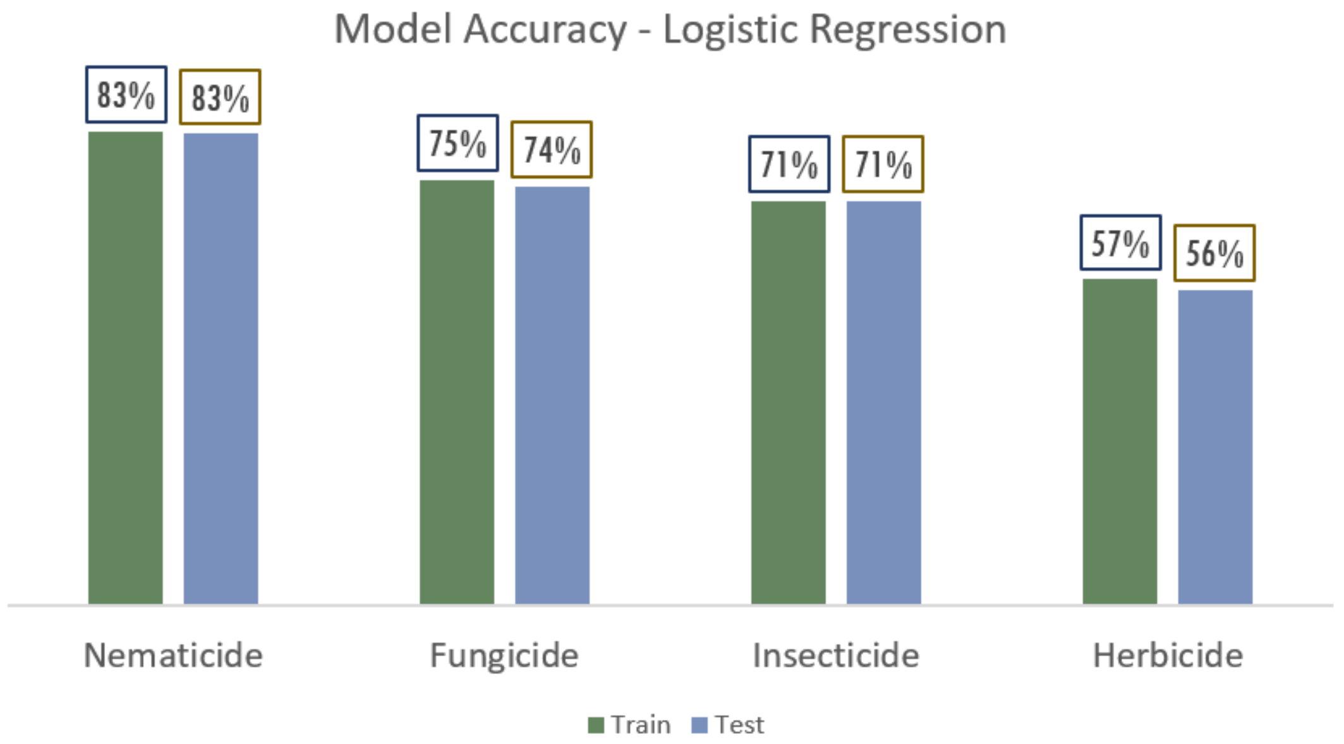


Figure 4. Model Accuracy for product type to be purchased

Figure 4 gives the accuracy of logistic regression in predicting whether a farmer would purchase the corresponding product or not. The models have good accuracy for predicting purchase of Nematicide, Fungicide and Insecticide. This information can be used by sales representatives to provide only the most relevant products to farmers and make better decisions on the amount of product to be produced.

Conclusions

Sales representatives want to know which customers are more likely to purchase their products and which specific products are they interested in buying.

The first model is used to find a segment of customers who have a high probability of purchasing the company's product. The second model is used to find out which products the customer is highly interested in buying. These segmentation in customer and product preference improve sales performance.

Acknowledgements

We thank Professor Matthew Lanham for his constant guidance on this project.