Tayko Software Cataloger

Predictive Modeling

Objective

Develop a model for classifying a customer as a purchaser or non-purchaser by implementing the following steps:

1. Partitioning the data into a training set (800 records), validation set (700 records), and test set (500 records).

2. Run logistic regression with L2 penalty, using method Logistic Regression CV, to select the best subset of variables.

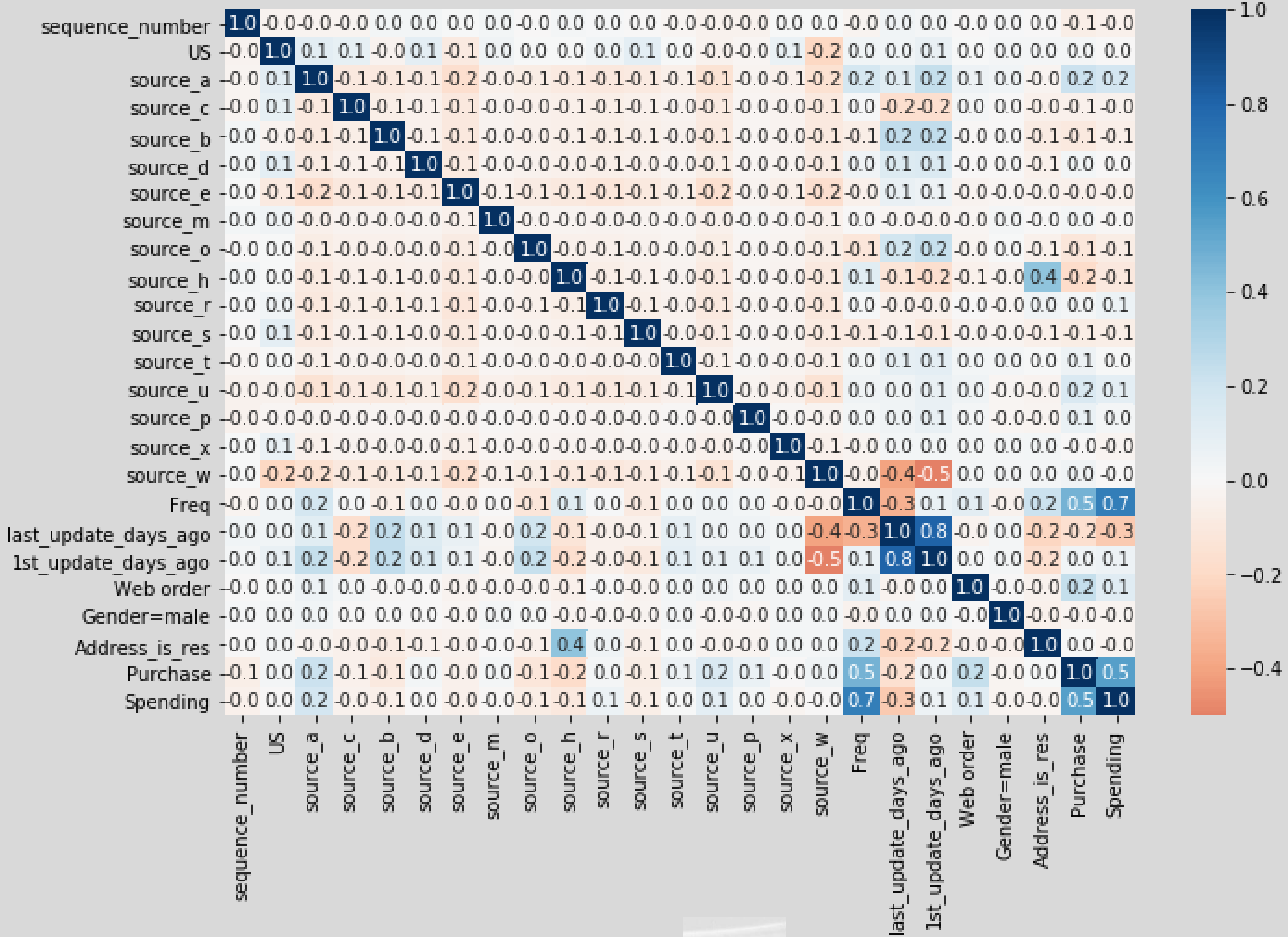
Exploration

Observation: Dataset has 25 variables which are mostly categorical in nature

Strategy: Correlation Matrix was formulated to deal with multicollinearity. From the matrix, we concluded:

Result: Spending and Freq, First update days ago and Last update days ago are highly correlated

Final Decision: Eliminate Spending and Last update days ago from our predictor list and Sequence number is just an id so it can be eliminated as well



Best Variables or Predictors

- ***Source A***
- ***Source C***
- ***Source D***
- ***Source H***
- ***Source P***
- ***Source R***
- ***Source U***
- ***Source X***
- ***Web Order***
- ***Frequency***
- ***First Update Days Ago***

Process of Variable Selection

1. Split the data into in the following sets:
   Training (40% = 800 records),
   Validation (35% = 700 records)
   Test (25% = 500 records)

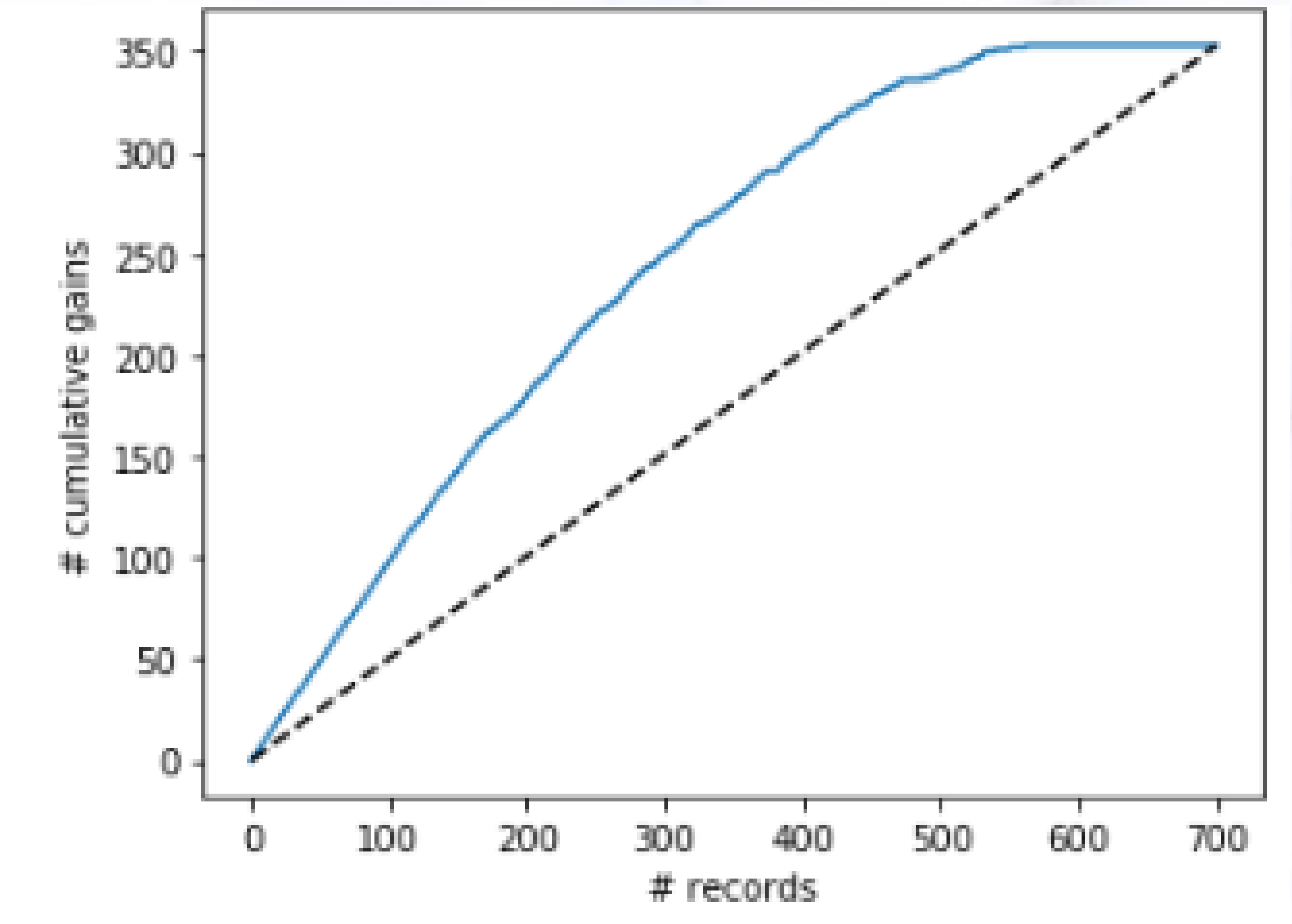
2. Techniques used for variable selection:
   Forward Selection
   Step Wise Selection

Model Comparison

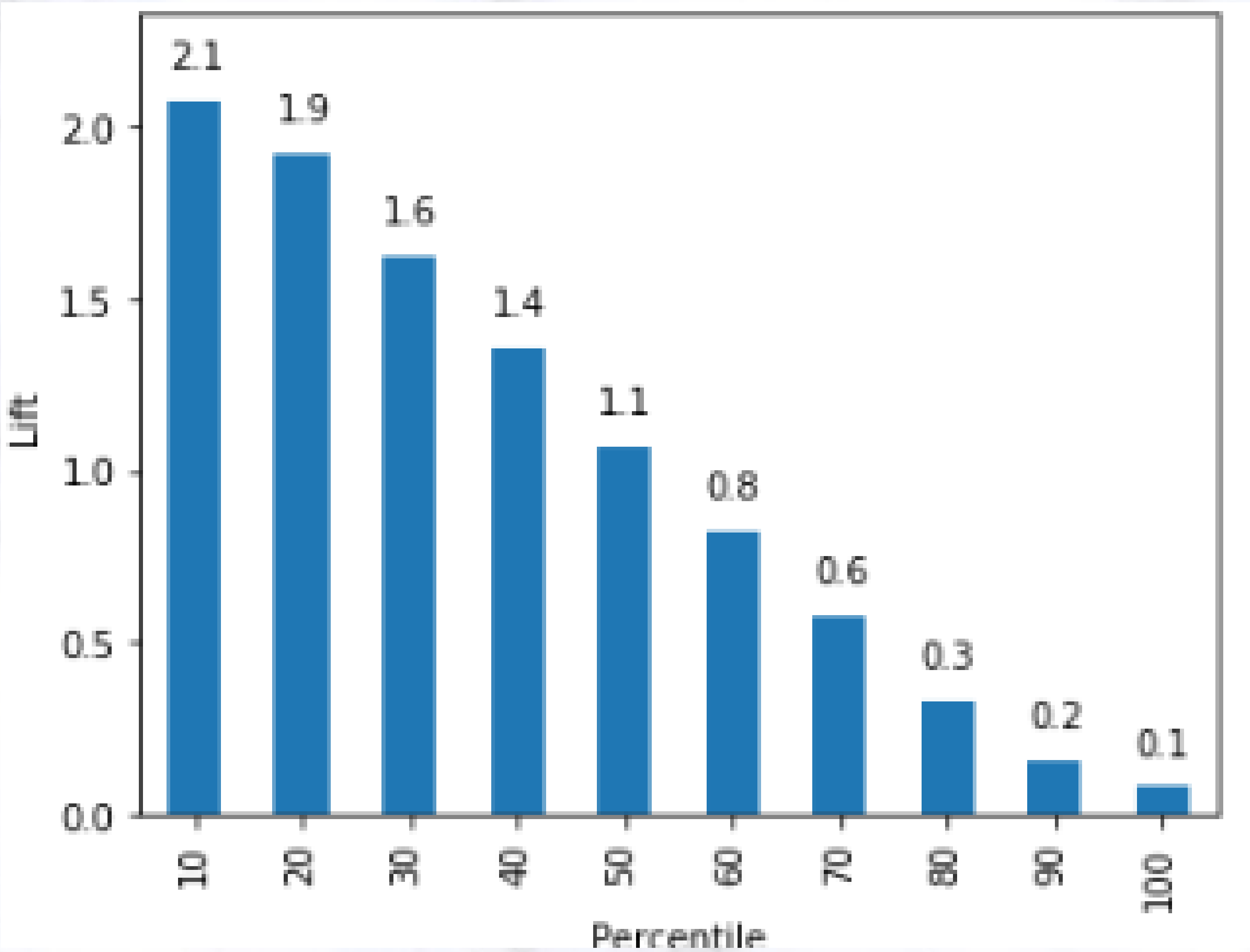
Variables with low probability of prediction	Before selecting Variables	After selecting Variables
	US, Source C, M,O,H, associated, First updated days ago, male	First Updated Days ago, Source C & Source H
Model Accuracy	78.86 %	78.00 %
AIC score	944	953

Conclusion: Our model accuracy before & after selecting variables is the same. It means that we are not losing any data by selecting the variables which is good for our model. Hence, we are reducing running time and saving our memory by using variable selection technique.

Final Result



Gains Chart- The blue line shows that our model is superior compared the original model represented by black dotted line. Greater the gap between the two, better the model we have.



Lift Chart- The first decile give us a lift by 2.1 as compared to a random selection. Since, it has the Staircase Effect i.e. bars descend in order from left to right, so we can go ahead with the model.