2020

An Analysis of the Success of Farmers Markets in Kentucky Using Logistic Regression and Support Vector Machines

Jeron Russell
Western Kentucky University, jeron.russell927@topper.wku.edu

AN ANALYSIS OF THE SUCCESS OF FARMERS MARKETS IN KENTUCKY USING LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINES

A Capstone Project Presented in Partial Fulfillment
of the Requirements for the Degree Bachelor of Science
with Mahurin Honors College Graduate Distinction
at Western Kentucky University

By

Jeron S. Russell

May 2020

*****

CE/T Committee:

Dr. David Zimmer, Chair

Dr. Melanie Autin

Dr. Dennis Wilson

ABSTRACT

The purpose of this research is to look at the relationship that market-specific, economic, and demographic variables have with the success of farmers markets in Kentucky. It additionally seeks to build a tool for predicting farmers market success that could be used by policy makers to aid in decision-making processes concerning farmers markets. Logistic regression and Support Vector Machines (SVMs) are used on data acquired from the Kentucky Department of Agriculture and the American Community Survey in order to analyze the data in a traditional statistical approach as well as a machine learning approach. The results included an SVM model that had an accuracy of 83.3% in predicting farmers market success. Additionally, both methods produced models that found population size, number of vendors, and number of years the market has been established as important predictors for farmers market success.

Dedication Statement

I dedicate this thesis to my parents, Elie and Diane, who instilled a love for learning and farming, and to my brother, Ethan, who was always my selling buddy at our local farmers market growing up.

ACKNOWLEDGEMENTS

I am grateful for the lessons learned during this thesis experience. My thesis advisor, Dr. David Zimmer, taught me the art of uncovering and telling the stories that data are concealing in their raw form, and he always had an open door for all of my quandaries. A special thanks to Dr. Melanie Autin, my second reader, who not only provided me with invaluable feedback for the completion of this thesis, but also opened up the world of statistics to me while offering constant support and encouragement throughout my collegiate experience. Finally, without the amazing professors in the WKU Department of Economics and Department of Mathematics constantly feeding my curiosity to learn, I would never have discovered my passion for analytics.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Farmers markets have been touted by proponents as mechanisms that encourage economic growth at a community level. Various studies using IMPLAN models have shown that farmers markets do tend to have a net positive economic impact (Otto & Varner, 2005; Cummings, Kora & Murray, 1998; Henneberry et al., 2008; Hughes et al., 2008; Hughes & Isengildina-Mass, 2015; Yosick, 2008). However, limited quantitative research that examines what makes an individual market successful has been done. This paper explores what variables may be influential in predicting farmers markets' success in Kentucky. Empirical results suggest that a mix of market specific variables and demographic variables can create simple models with good predictability power for farmers markets' success with both traditional and machine learning methodologies. This research makes no claims about causal relationships. When 'relationship' or 'effect' is used in interpretation it merely describes the association between the independent and dependent variable.

The research presented in this thesis shows evidence contrary to findings by Schmitz (2008) which found household median income to be correlated with farmers market gross market sales. This thesis, which uses gross market sales as a proxy for farmers market success, finds household median income to not be a significant predictor in any model. Research by Bonanno, Berning, and Etemaadnia (2017) also found median income to be insignificant. This additionally agrees with research done by Wolf, Spittler, and Ahern (2005), which found income of farmers market consumers to be no different than that of other consumers, but it disagrees with work by Gumirakiza, Curtis, and Bosworth (2014), which found high-income females to be the most likely to attend farmers markets. Similarly to Wolf's et al. (2005) finding of no difference in employment level between farmers market consumers and other consumers, my research suggests unemployment is not a significant predictor to farmers market success. Bonanno et al. (2017) found that many demographic and economic variables, including Supplemental Nutrition Assistance Program (SNAP) participation, were significant predictors of the number of farmers markets in a zip code across six New England states. Empirical evidence from my research finds similar results to Bonanno et al. (2017) with respect to population size and median age showing predictive power for market success, but it finds contrary evidence to a majority of the demographic and economic variables, including SNAP participation, which Bonanno found to be significant. Perhaps Kentucky markets perform differently from New England markets, or perhaps number of markets in a zip code and success of a market are not directly correlated.

This thesis uses logistic regression and Support Vector Machines (SVM) to analyze the ability of economic and demographic variables to predict farmers market success in Kentucky. Findings show that the number of vendors at a market offers the most information about the success of a farmers market. Furthermore, both methods indicate that population size surrounding the market is an important predictor as is the number of years the market has been established. Median age of people near the market is found to be a good predictor for the SVM model but not for logistic regression. No economic variables or any other demographic variables are found to be good predictors of farmers market success. Additionally, results show that larger population size, more vendors, and higher rates of market longevity are associated with a higher likelihood of a farmers market being successful. Furthermore, the SVM model created is able to predict the success of a farmers market correctly 83.3% of the time on unseen data. Another interesting finding is that the percentage of SNAP participants was not found to be useful in either model. This finding indirectly offers support that, perhaps, the Kentucky Double Dollars program is doing a good job of incentivizing SNAP participants to attend farmers markets which often have higher prices than grocery stores. If SNAP participants are able to attend markets as often as non-SNAP participants, it is believable that SNAP participation shouldn't offer any information about farmers' market success. This thesis also investigates the similarities and differences in methodology between traditional logistic regression used in economics and statistics and more modern Support Vector Machine models that have gained popularity with machine learning.

From a policy standpoint, a tool that is able to predict the successfulness of a farmers market may aid in the decision-making process when distributing farmers market specific grants. Knowing in advance what markets are likely to fail versus those that appear to be struggling despite having all the characteristics for success could enable grant money to be used more effectively. Additionally, knowing good predictors of farmers market success can aid in picking locations for new farmers markets. For example, results from this research show empirical evidence that indicates considering population size near the market may be important. The results of this research lead to theorizing that farmers market success might be better predicted by market specific traits. If future research supports this theory, it would allow farmers markets' managers to be better informed on the power they have to ensure farmers market success.

CHAPTER TWO: LITERATURE REVIEW

Much of economic farmers market research is concerned with the economic impact that farmers markets have on the economy. A 2002 paper by Brown takes a comprehensive look at research done between 1940 and 2000. She concluded that up until 2000 the literature addressing the economic impact is “meager” and that, overall, the limited attempts to quantify farmers markets’ effects have still left the best methods for analysis unclear. More current research, whose methods will be discussed later, done by Hughes, Brown, Stacy, and McConnell (2008) calls Brown’s (2002) claim that there had been no effective method for evaluating economic impacts an erroneous one, but the research methods cited as evidence against Brown’s (2002) claim all come after the publication of Brown’s article; thus this would not devaluate Brown’s claims up until 2000.

Brown (2002) actually defends the possibility that a quantitative method can be found but addresses the shortcomings up until her paper was published. Brown points to an issue with differing definitions of who gets included in direct market studies that makes it nearly impossible to compare results from study to study. Other key issues that Brown (2002) mentions include numbers being reported as averages that do not take into account differing vendors or items sold, as well as the data being provided tending to substantially underestimate actual values. Many of the issues can be solved through improved data collection methods and more clarity in research methods. Note that understated or limited data may lead to questions about the validity of research findings, but it does not discredit the methods that were used.

Though research regarding farmers markets' economic impact is still limited, new methods have become popular for analyzing markets' effects. Chiefly among these methods are the use of IMPLAN models. Though this method of analyzing economic impact by using an input-output model was developed by the USDA in the 1970s, it didn't become a popular choice for farmers market research until the early 2000s. There have been a multitude of IMPLAN models performed on farmers markets across the U.S. and Canada. Otto and Varner (2005) used an IMPLAN model, and their results indicated that $31.5 million were generated directly or indirectly by farmers markets gross profit in Iowa. Cummings, Kora, and Murray (1998) also used an IMPLAN model and estimated that $1 billion in secondary effects can be attributed to farmers markets in Ontario. Another study looking at farmers markets in Oklahoma indicated that the gross profit of $3.3 million generated from the markets have a total economic impact of $7.8 million on Oklahoma's economy (Henneberry et al. 2008). Assuming accurate data from these studies that used voluntary surveys, results indicate that farmers markets have a varying impact based on their location, but all showed a positive economic impact.

Hughes et al. (2008) and Hughes and Isengildina-Mass (2015) took issue with the methods the studies above used. While they had no issue with the IMPLAN model, which they also used, they believed that the studies overstate economic impact because gross numbers were used and opportunity cost was not accounted for. Hughes et al. (2008; 2015) conducted two separate studies on the farmers markets of West Virginia and South Carolina. An IMPLAN model was used to calculate gross economic impact and then estimated for opportunity cost were developed. The study then took the gross impact and subtracted it from the opportunity cost to calculate net farmers market economic impact.

The net economic impacts were smaller than other studies, but still indicated a positive economic impact. The study claims that biased policy makers favor farmers markets research that uses gross numbers and thus show a higher net impact, when, in actuality, an opportunity cost framework tells a more realistic story.

While IMPLAN seems to be the most popular choice for analyzing economic impacts, there are studies that have taken a more statistical approach when analyzing farmers markets. A Master's thesis by Schmitz (2008) did comprehensive analysis on markets' performance based upon their urban classification. A Kruskal-Wallis statistical method was used for analysis on farmers markets in Kentucky. The results indicated that farmers markets located in micropolitan areas tended to have the longest lasting farmers markets compared to metropolitan and rural markets. The research further showed that metropolitan areas had the highest gross market sales followed by micropolitan and then rural markets. Schmitz's (2008) research also noted that there is a correlation between household median income and gross market sales.

Researchers have disagreed on the size of the economic impact that farmers markets have, but the vast majority of research using IMPLAN modeling has indicated that farmers markets do have a positive economic impact. This is not to say that every farmers market has the same impact, nor that the impact will be constant over time; not all markets are created equal. Examining what traits are able to explain the success of farmers market is a logical next step in order to understand how to get the most economic benefit out of farmers markets. Uncovering these variables is instrumental in helping new markets be successful and for government agencies to be better informed when awarding grant money to markets.

Limited quantitative research that looks at what economic factors make a farmers market successful exists. An exception to this is the work done by Bonanno et al. (2017). Research by Bonanno et al. (2017) includes data analysis at the zip-code level for six New England states. A probit model is used to model the relation between demographic and economic data and the number of farmers markets that are in a zip code. The results of their model indicated population size, households with children, race, percentage of SNAP participants, household size, education level, and median age are all significant in predicting the number of farmers markets in the zip code. It also found that household median income was not a good predictor of the number of markets. Following the line of thinking that the traits that can significantly predict the number of markets in a zip code should also correlate to the success of a market, many of the same demographic and economic variables were used in the research presented in this paper. The assumption is also made that if there is a common demographic of a farmers market consumer, then areas high in that demographic should lead to a more successful market. Thus, knowing the profile of a 'typical' farmers market customer could be beneficial.

Research is mixed and limited primarily to survey data with respect to the profile of a farmers market consumer. Research by Wolf et al. (2005), which looked at 336 produce consumer responses in San Louis Obispo County, California, indicates that age levels, income levels, and employment status are similar between farmers market shoppers and farmers market non-shoppers. This implies they are not unique traits to farmers market consumers and thus shouldn't really have an effect on markets performance. This is contrary to results by Bonanno et al. (2017) which indicated that age is a significant predictor in number of farmers markets in a zip code. Research by Gumirakiza et al. (2014), which analyzes survey responses from 1,488 consumers across Nevada and Utah, found that farmers market attendees tend to be female with higher income levels. This finding contradicts Wolf's et al. (2005) and Bonanno's et al. (2017) findings that income is not important to a farmers markets' consumer profile nor an explanation for number of markets in a zip code. The research presented in this thesis helps add information to the sometimes conflicting results currently in literature.

CHAPTER THREE: DATA

The data used in this research consists of 2017 farmers market data acquired from the Kentucky Department of Agriculture (KDA) and 2017 5-year estimates of economic and demographic data acquired from the American Community Survey. Data acquired from the KDA included the number of vendors at a market, the years that the market has been established, and the gross profit of the market. As noted by Brown, farmers market profit is often underreported. Additionally, when numbers were provided by the KDA, they explained that numbers were more than likely underreported since the data was given voluntarily. Due to the possible unreliability of the exact profit numbers, this research uses the profit as a proxy for the success of a market. A "successful" market was designated as any market that performs better than the median of gross profit across all 119 markets. All markets that have gross profit at or below the median were designated "poor-performing." This resulted in 59 markets being categorized as successful and 60 as unsuccessful.[1]

Each markets' zip code was retrieved from the KDA. The zip code was used to find the demographic and economic data from the American Community Survey. Data retrieved for each represented zip code were household median income, median age, population size, percentage of people with a bachelors, percentage of people that are white, percentage of people that are on SNAP, and unemployment rate. The dependent variable is market success classification (success or failure), and all other variables are candidates for independent variables. Table 3.1 displays

[1] In this study, gross profit from the market was used as a proxy for success. It followed a simple idea that markets that brought in more money would be able to impact more consumers and provide a higher net economic impact. There is nothing inherently special about the median being used as a cutoff, as it is only a proxy. The appendix provides a logistic regression model with an alternative definition of success as its dependent variable.

summary statistics for each independent variable. The choice of variables was based off previously mentioned work by Bonanno et al. (2017), Gumirikiza et al. (2014), and Wolf et al. (2005).

	Mean	Median	Minimum	Maximum
Population	16,988.86	13,980.00	147.00	61,471.00
Unemployment(%)	7.21	6.50	0.00	26.50
Bachelors (%)	20.58	16.20	7.60	67.30
White (%)	91.54	93.80	42.60	100.00
Household Median Income ($)	44,523.85	40,347.00	17,372.00	135,190.00
SNAP(%)	18.93	17.10	1.70	55.10
Median-Age (Years)	40.19	40.90	26.90	59.20
Vendors	17.83	13.50	2.00	86.00
Years Established	18.18	15.00	1.00	58.00

Table 3.1: Summary Statistics of Independent Variables

The variables are not all normally distributed without transformation. Logistic regression is the first model that is used to analyze the data. Logistic regression does not require the assumption of normality, and thus data does not undergo any transformations. When comparing across different machine learning models, it is important for data to be normal and to be on a similar scale. When analyzing machine learning models' performance, and ultimately deciding on using a support vector machine, the data was first normalized and scaled by using the MinMaxScaler() function in python (Pedregosa *et al.*). This function takes

$$\frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

in order to transform data into a range between 0 and 1.

CHAPTER FOUR: A TRADITIONAL STATISTICS APPROACH WITH LOGISTIC REGRESSION

4.1 Model and Analysis

Logistic regression is a commonly used model across statistics and economics for modeling events with a binary outcome. Often the value 1 is equated with the success of an event, while 0 is equated with failure of the event to occur. In the case of farmers markets, 1 is assigned to the markets deemed successful, while all other markets are assigned a value of 0. Throughout this thesis, markets that receive a 0 may be referred to as failures, poor performing, or unsuccessful farmers markets. These all refer to markets that were not categorized as a successful market. Logistic regression allows for independent variables that are both continuous and categorical, though analysis in this paper only has continuous predictors. Logistic regression uses odds,

$$\text{odds} = \frac{\pi}{1-\pi} = \frac{P(y=1)}{P(y=0)}, \tag{4.1}$$

and log odds,

$$\text{log-odds} = \ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{P(y=1)}{P(y=0)}\right), \tag{4.2}$$

where π is the true probability of a successful response, in order to estimate the response variable. With k predictor variables the model fit is

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{4.3}$$

where x_i represents the i^{th} predictor variable (Agresti, 2007).

This model illustrates that the coefficients and predictor variables are a linear combination of the log-odds of an event occurring, i.e., a one-unit increase in x_i leads to an additive change in log-odds of the event by β_i. The log-odds of an event occurring has little meaning to most people, so an arithmetic reorganization of (4.3) yields

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}. \tag{4.4}$$

From equation 4.4, it can be derived that a one-unit increase in x_i has a multiplicative effect of e^{β_i}, the odds ratio, on the odds of the event occurring. Furthermore, if $e^{\beta_i} > 1$, the probability of success, π, increases as x_i increases. If $e^{\beta_i} < 1$, then π decreases as x_i increases. If $e^{\beta_i} = 1$ then π does not change as x_i increases. This assumes that all other variables are held constant (Agresti, 2007).

In many traditional applications of logistic regression, especially in economics, all of the data available is often used to fit the model, and rarely is data split into a training set and a test set. Though it is possible for logistic regression models to do this, and is common practice in machine learning, this thesis refrains from doing so when using logistic regression. This thesis, while primarily focused on exploring farmers market success, also seeks to address the differences between traditional and machine-learning analysis, and thus for methodological comparisons the data has not been split for the logistic regression model. Additionally, logistic regression is not analyzed as a classifier which would set a cut-off probability for success and failure though this is often done in machine learning.

4.1.1 Hosmer and Lemeshow Goodness-of-Fit Test

The Hosmer and Lemeshow Goodness-of-Fit (HL) Test is a statistical test that addresses the goodness-of-fit of a model. The test has the null hypothesis (H_0) that the model is a good fit and the alternative hypothesis (H_a) that the model is not a good fit.

The HL Test tests the hypothesis by looking at observed versus predicted values across different groupings. Let g_i represent the i^{th} group into which the observations are split into. Suppose $i \in (1, 2, \ldots, 10)$ then g_1 consists of the observations with the lowest 10% predicted probabilities, and g_2 consists of the 10% of observations from the sample whose predicted probabilities are the next smallest, etc. For instance, the farmers market data in this thesis includes 119 farmers markets. Model 3, presented in the following section, has g_1 consisting of 12 (approximately 10% of the observations) farmers markets, and those 12 farmers markets have an expected value of 0.85 for success. That is, out of the 12 markets, it is expected that 0.85 of them are successful farmers markets. In actuality, 1 market out of the 12 was successful. Additionally, the expected number of farmers markets that are unsuccessful is calculated as 11.15 while the observed number is 11. Similarly, g_2 has 12 farmers markets, and its expected value for the number of successful markets is 1.48. The observed value is 1. It has an expected value of unsuccessful markets of 10.52 and an observed amount of 11. This continues for all 10 groups. The HL Test then compares these expected versus observed values across all 10 groups. The closer the observed and expected values are, the better the fit is determined to be. The test statistic is calculated as

$$H = \sum \frac{(\text{Observed Successes} - \text{Expected Successes})^2}{\text{Expected Successes}} + \frac{(\text{Observed Failures} - \text{Expected Failures})^2}{\text{Expected Failures}} \qquad (4.5)$$

and has a chi-square distribution with $i-2$ degrees of freedom. The summation includes all groups; in this case 10.

There are several different tests that address the goodness of fit of a logistic model, but the HL Test will be the primary method in assessing model fit in this paper. Additionally, AIC values will be presented.

4.1.2 Receiving Operating Characteristic Curve

One way to analyze the predictability power of a model is using a receiving operating characteristic (ROC) curve. On the y-axis, ROC curves plot sensitivity, also known as the true positive rate, where:

$$\text{Sensitivity} = \frac{\text{\# True Positives}}{\text{\# Positive}} = \frac{\text{\# True Positives}}{\text{\# True Positives+ \# False Negatives}}. \qquad (4.6)$$

On the x-axis, Roc curves plot 1-specificity, also known as the false positive rate, where:

$$\text{Specificity} = \frac{\text{\# True Negatives}}{\text{\# Negatives}} = \frac{\text{\# True Negatives}}{\text{\# True Negatives+ \# False Positives}}. \qquad (4.7)$$

1-Specificity is the probability of predicting a real negative as positive.

Logistic regression provides estimated probabilities that an event might occur, not strict classifications of success and failure. In order to classify an event, a cutoff point must be designated. If 0.5 is used as a cut off value, then if an estimated probability is at or above 0.5 it will be classified as a positive event; if it is below 0.5, it will be classified as a negative event. ROC curves use all possible cutoff values between 0 and 1 to plot sensitivity and 1-specifictiy (Agresti, 2007). Analyzing the Area Under The Curve (AUC) gives information about the predictability power of the logistic regression model.

4.2 Logistic Regression with Farmers Market Success

Table 4.1 displays the estimated coefficients of independent variables for three logistic regression models using market classification (success or failure) as the dependent variable.

	Model 1	**Model 2**	**Model 3**
Population	.000050	.000078	.000075
	(2.6836)	(8.7862)***	(9.2277)***
Unemployment %	-0.0700	-0.1022	--
	(0.5347)	(1.5935)	
Bachelor's %	-0.0130	0.0171	--
	(0.1548)	(0.4717)	
White %	-0.0529	-0.0272	--
	(1.7340)	(0.6692)	
HMI ($)	1.485x10^{-6}	-0.00003	--
	(0.0023)	(1.3329)	
SNAP %	-0.0373	-0.0169	--
	(0.8788)	(0.2355)	
Median Age	-0.0654	-0.0417	--
	(0.5548)	(0.5652)	
Vendors	0.1501	--	0.1409
	(19.9978)***		(19.1034)***
Years Established	0.0328	--	0.0354
	(2.2149)		(2.9832)*
Hosmer-Lemeshow Goodness-of-Fit Test (p-value)	0.0473	0.3439	0.9792
AIC	118.50	149.39	110.95

Table 4.1: Model Coefficients, AIC and HL Test

Note: Values in parentheses indicate the chi-square test statistic. *, **, *** indicates significance at the 0.1, 0.05 and 0.01 level, respectively.

4.2.1 Model 1 (All Variables)

The estimated coefficients for model 1 are presented in table 4.1, but analysis for these variables will be skipped due to the HL Test's indication that this is not an appropriate model. Furthermore, only number of vendors was found to be a statistically significant predictor of farmers market success while all other predictors are in the model. The model has an AIC value of 118.50.

4.2.2 Model 2 (Demographic and Economic Variables)

Model 2 excludes market-specific variables in order to analyze the ability of a model with only economic and demographic variables to predict market success. The HL Test indicates that this model is a good fit, but the only predictor that was found to be significant while all other demographic and economic variables are in the model is population size. It also resulted in the highest AIC value at 149.39. Furthermore, population size is the only variable that remains when doing forward-selection procedures. This indicates that population is the best predictor of market success, while the other variables do not offer much information. As a result, only the interpretation of population is included for this model.

For every 1000-person increase in population, it is expected that the odds of a farmers market being a success will increase by 8.11% assuming that all other demographic and economic variables are held constant.

4.2.3 Model 3 (Best Model)

Model 3, chosen through a forward-selection procedure, resulted in a good fit according to the HL Test. It also has the lowest AIC value at 110.95. This model includes population size, number of vendors, and number of years the market has been established as independent variables. All three predictors are significant at the 0.1 significance level, while population size and number of vendors are also significant at the 0.01 significance level.

To reiterate, the purpose of this thesis is not to seek a causal relationship between market success and certain predictors. It is more exploratory in nature. It is important to notice that when predictors' "effects" on the response variable are analyzed, it is merely examining the relationship between market success and the predictors. Also, the scope of the inference is limited to farmers markets in Kentucky, not farmers markets in general. The resulting logistic regression model for model 3 is:

$$\ln\left(\frac{\text{P(Farmers }\widehat{\text{market}}\text{ success)}}{\text{P(1}-\text{Farmers }\widehat{\text{market}}\text{ success)}}\right) = -0.0569 + 0.000075(Pop.) + 0.1409(Vendors) + 0.0354(YrsEst). \tag{4.8}$$

Using the fitted model in (5.8), we can interpret the estimated coefficients for each predictor variable in the logistic regression model. These interpretations are summarized in table 4.2.

For every 1000-person increase in population size, it is expected that the odds of a farmers market being a success will increase by 7.79%, assuming that the number of years the market has been established and the number of vendors are held constant. Note that this is similar to model 2's findings.

For each additional vendor, it is expected that the odds of a farmers market being a success will increase by 15.13%, assuming that the number of years the market has been established and population size are held constant.

For each additional year the market has been established, it is expected that the odds of a farmers market being a success will increase by 3.60%, assuming number of vendors and population size are held constant.

Variable	Effect of Odds of Market Being Successful
Population Size	7.79% Higher
Number of Vendors	15.13% Higher
Number of Years the Market Has Been Established	3.60% Higher

Table 4.2: Effect of Odds of Market Success

Effects are for a 1000-person increase in population and a one unit increase for other variables

By standardizing the coefficients, it can readily be seen which variables have the largest impact on the model. Table 4.3 shows the standardized coefficients of the predictor variables. Standardizing shows that number of vendors has the largest impact, followed by population size and, lastly, the number of years the market has been established.

Variable	Standardized Coefficients
Population Size	0.5529
Number of Vendors	1.0955
Number of Years the Market Has Been Established	0.2470

Table 4.3: Standardized Beta Coefficients

4.2.4 Assessing Predictive Power of Model 3 with an ROC Curve

A completely random model that offers no predictability power is represented by the straight line passing from (0,0) to (1,1) in Figure 4.1. The other curve represents the ROC curve of Model 3. As predictability power increases, the curve moves closer to the top left corner of the graph. The Area Under the Curve (AUC) is 0.8777. As a rule of thumb, a model with AUC of greater than 0.8 can be seen as having good predictability power. The AUC tells us that if two markets were picked at random, one of them being a successful market and the other being a poor performing market, the successful market would be ranked higher; that is, the probability of success would be higher than the poor performing market 87.77% of the time. This would be considered a model with good predictability power based on the analysis of the ROC curve.

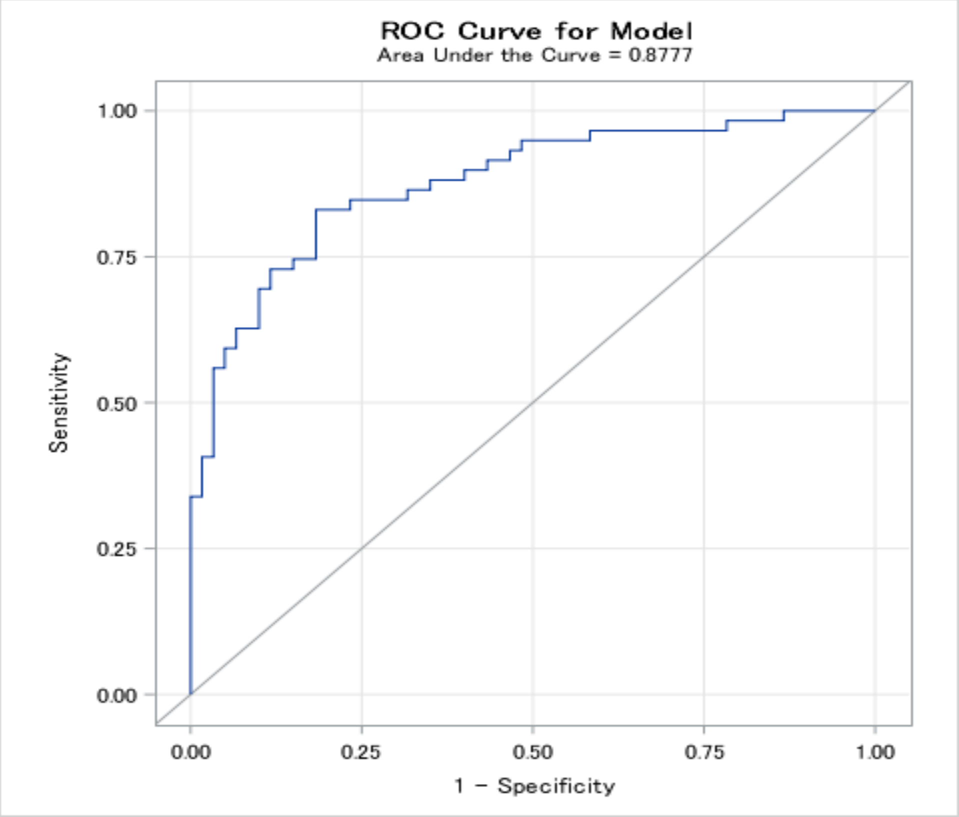

Figure 4.1: ROC Curve for Model 3

CHAPTER FIVE: A MACHINE LEARNING APPROACH WITH SUPPORT VECTOR MACHINES

5.1 Support Vector Machine Analysis

The Support Vector Machine was chosen over other classification machine learning models based on its having the highest mean accuracy and lowest standard deviation of the accuracies from 5-fold cross-validation with all features included in the model. This was measured against K-NN, Naïve Bayes, Random Forest, and Decision Tree classifiers. To stay consistent with machine learning terminology, the variables that were discussed in Chapter 4 as independent variables will be referred to as features in Chapter 5.

In this paper, 5-fold cross-validation is also used to optimize hyper parameters in the Support Vector Machine model. It will further be used to aid in the analysis of the performance of the model alongside the accuracy resulting from the test set that will be used for final evaluation of the model.

This paper offers a brief description of 5-fold cross-validation for those not familiar with machine learning processes. 5-fold cross-validation takes the training dataset and divides it into 5 groups. Each group is withheld individually while the model is trained on the remainder of the training set data and then tested on the group that was withheld. Each group becomes a test set, and there are thus 5-folds. The illustration in Figure 5.1 is helpful in understanding the process.

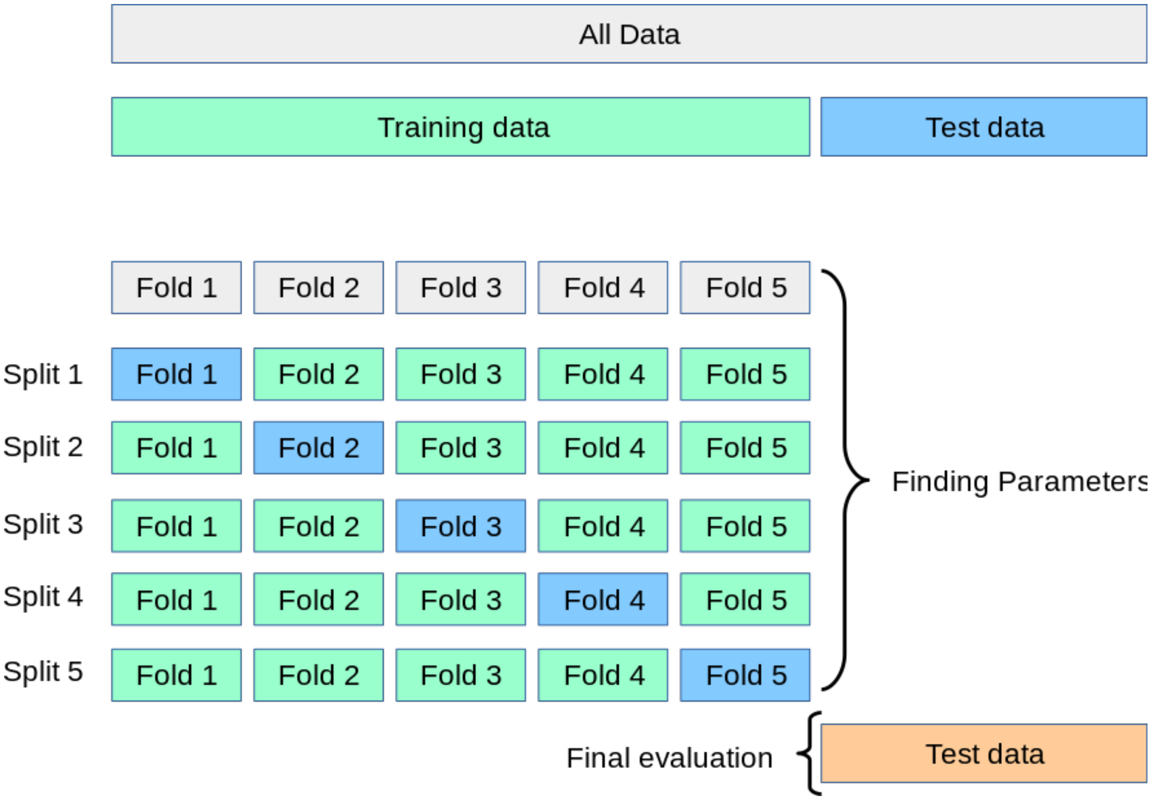

Figure 5.1: 5-Fold Cross-validation (Pedregosa *et al.*).

Support vector machines became prominent after being introduced by Cortes and Vapnik in 1992. SVM became very popular and very useful because of its ability to take data that is not initially linearly separable and make it linearly separable without needing too much computational computing power. This is accomplished by using a kernel function. A kernel function is defined as a function that takes, as its inputs, vectors in the original space, and returns the dot product of the vectors in the feature space. More formally, let X be represented by an $n \times n$ kernel matrix of pairwise similarity comparisons where entries (i, j) are defined by the kernel function $k(x_i, x_j)$. If we have data vectors $x, x' \in X$ and a map $\varphi: X \rightarrow R^n$ then

$$k(x, x') =< \varphi(x), \varphi\ (x') > \tag{5.1}$$

is a kernel function.

For the rbf kernel,

$$k(x, x') = exp(-\gamma||x - x'||^2), \tag{5.2}$$

where $\gamma = \frac{1}{2\sigma^2}$.

For the linear kernel,

$$k(x, x') = < x, x' > \tag{5.3}$$

(Christianini and Schölkopf 2002). Many other kernels exist, but linear and rbf kernels are the two that are used in this research.

The intuition behind SVMs can be more readily understood using Figure 5.2 as a graphical representation of how a linear kernel separates data. A linear kernel assumes that the data is already linearly separable. A hyper-parameter is used to determine the distance between the margins. A high value for C results in a smaller margin, while a lower value for C results in a larger margin. The implications of a larger and smaller margin are discussed in more detail in the following paragraphs. The line between the margins is the decision boundary that is used for classification decisions.

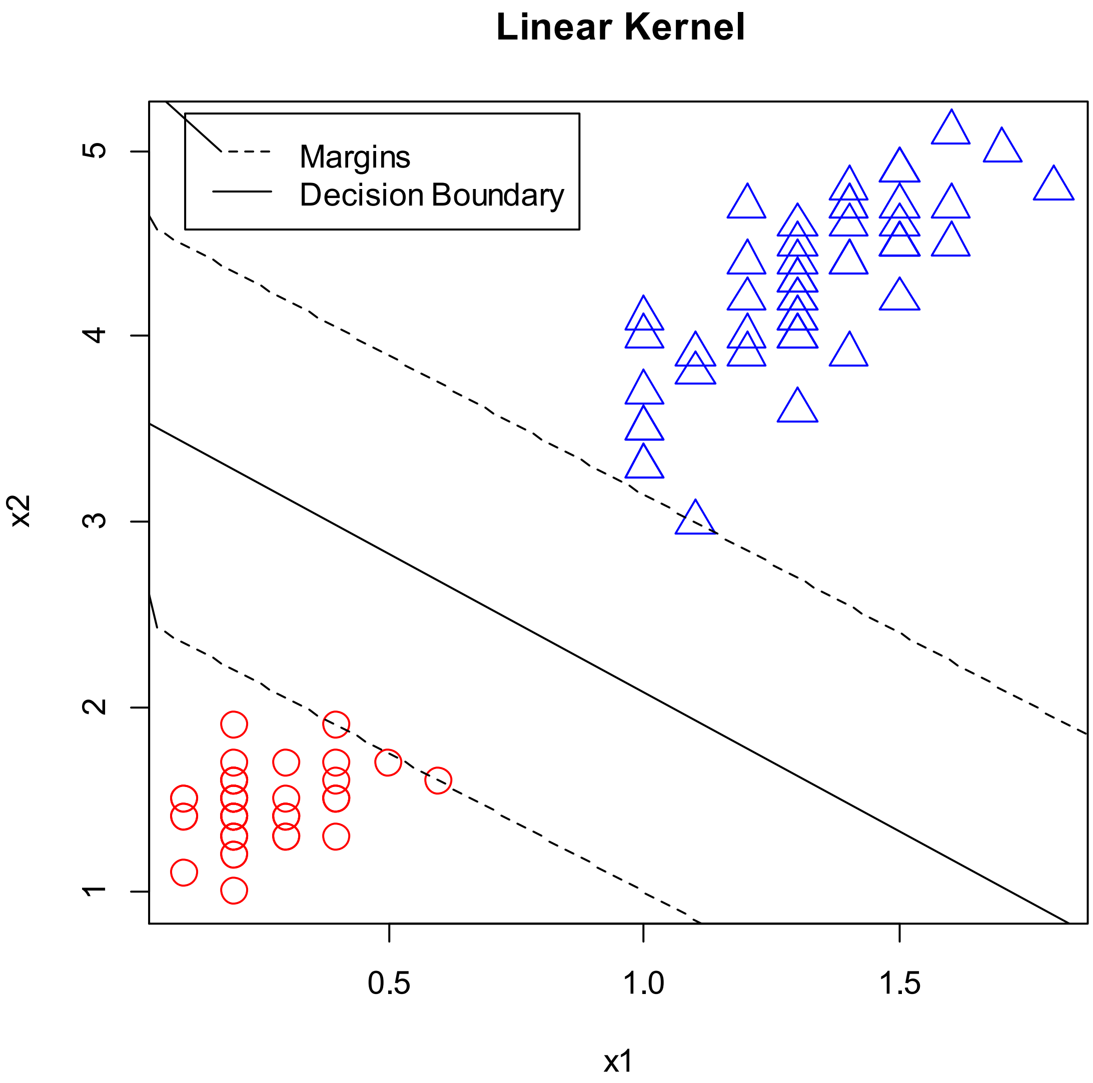

Figure 5.2: Linear Kernel

Figure 5.3 is a graphical representation of how a radial basis function kernel can take data that is not linearly separable in its original state and make it linearly separable in a higher dimension once the kernel is applied.

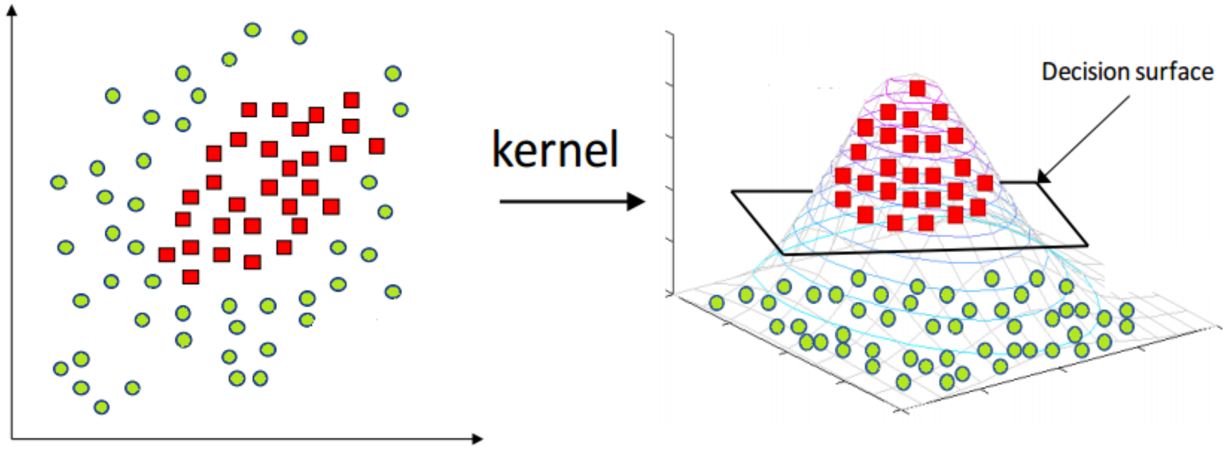

Figure 5.3 RBF Kernel (Zhang, 2018)

As mentioned, 5-fold cross-validation will be used to assess what hyper-parameters are to be used. Hyper-parameters include the type of kernel, in this case linear or rbf. For both models a C parameter must be set. C is a value that will decide how much to punish the model for misclassification on the training set. By way of adjusting the distance between the margins, a high C value will result in less misclassification in the training set, while a low C will result in more misclassifications in the training set. Gamma is a parameter used for the rbf kernel that tells the model the amount of influence a single training example should have. A small gamma tells the model a training example should have a far reach, while a large gamma tells the model a training example should have a close reach (Pedregosa *et al.*).

Three models are analyzed with the SVM algorithm. The first model includes all the features of the data set. The second model includes only demographic and economic variables. This is done in part to analyze how well the model will do without market-specific variables. The third model uses population size, median age, number of vendors, and the number of years the market has been established as features. These features were selected using an ANOVA F-Value method. This method checks each feature independently to see if the means of the features are different when grouped by binary classification of farmers market success. Through sklearns SelectKBest function, the ANOVA F-Value method was used, and the top four features F-value scores were kept in the model (Pedregosa *et al.*). This is a one-way ANOVA method for feature selection, and the F-value is calculated as

$$\text{F-Value} = \frac{\text{Variance of the group means}}{\text{Mean of the within group variances}}. \qquad (6.4)$$

5.2 Support Vector Machines with Farmers Market Success Results

5.2.1 Model 1 (All Variables)

Initially, farmers market data is separated into two groups. Group one is a training set, which is composed of 70% of the data. This is the data that the SVM uses to build its model. The training set is also the data that is used for 5-fold cross-validation. The other 30% of the data is kept separate from the rest of the data; this data is used to test the performance of the model on farmers market data that the model has not been trained on.

To begin the classification of farmers market success using a support vector machine, first an initial model is used to evaluate performance when all the features (household median income, percentage of population that is white, percentage of population that are SNAP recipients, percentage of the population with at least a bachelor's degree, median age of the population, population size, unemployment rate, number of vendors in the market, and number of years that the market has been established) are included in the model. The grid search function in Python, which is set to pick hyper-parameters that optimize the model for accuracy based on the mean of the 5-fold cross-validation accuracies, determined that the hyper-parameters should be kernel=rbf gamma=0.02, and C=1000.

The choice of the rbf kernel implies that the data is not easily linearly separable in its original input state, and, thus, an rbf kernel is used to transfer the data into a higher dimension feature space and make the data more easily linearly separable by a hyper-plane. The rbf kernel improves the accuracy of the model, but it also loses interpretability of the features in the model. Since the data has been transformed by the rbf kernel, the effect each feature has on the classification decision of farmers market success cannot be easily evaluated.

Table 5.1 displays the 5-fold cross-validation accuracies of the initial model that includes all features. The accuracy is simply the percent of data points that are classified correctly as a successful market or as a poor performing market under each fold, or division, of the data set. The mean of these accuracies is 0.782, and the standard deviation of the accuracies is 0.052. The low standard deviation is an indication of low variance in the model. Since the model has a fairly high accuracy with low variance, this provides evidence that the model will be able to fit well to data it has not seen and is not creating a model that is over-fitting the data. It is important to keep in mind that 5-fold cross-validation is performed on training data, while test data is still untouched by the model.

Fold	Accuracy
1	0.765
2	0.765
3	0.882
4	0.765
5	0.733

Table 5.1: 5-fold Cross-validation Accuracies for Model 1

Though 5-fold cross-validation is largely used for parameter selection, from the 5-fold cross-validation we would expect the model to classify new farmers markets correctly 78.2% of the time.

Now that the model has been trained and indications show that the model is not over-fitting the data, it is important to examine the model performance on the test data. This will give a further indication of how well the model is at classifying new data points as well as an indication of whether or not over-fitting is present.

	Predicted Market Failure	Predicted Market Success
Actual Market Failure	17	4
Actual Market Success	3	12

Table 5.2: Confusion Matrix for Model 1

The confusion matrix displayed in Table 5.2 shows that there are 17 true negatives, 4 false positives, 3 false negatives and 12 true positives. The accuracy of the model on the test data is calculated as

$$\frac{17+12}{17+4+3+12} = \frac{29}{36} = 0.806.$$

This indicates that the model is expected to classify 80.6% of new farmers market correctly.

The accuracies from the 5-fold cross-validation and the accuracies from the test set illustrate that the model that includes all of the features produces consistent performance on data points it has not seen before.

5.2.2 Model 2 (Demographic and Economic Variables)

Again, data is separated into two groups, a training set comprised of 70% of the data and a test set of 30% of the data. It is important to note that the training and test sets are made up of the same observations as in the first model.

This model excludes the features that were farmers market specific. That is, it excludes the numbers of vendors at the market and the number of years the market has been established. The remaining features are all economic and demographic information in each farmers market's respective zip code. Using the grid search function with the remaining features results in hyper-parameters of kernel=rbf, gamma=0.02, and C=100.

Once again, an rbf kernel is chosen to optimize the model based on the mean of the 5-fold cross-validation. This implies that this model, like Model 1, is more easily separated by a hyper-plane in a higher dimension than in its original space. Because of this, it is still not appropriate to analyze the weights of the features on the classification decision.

Table 5.3 displays the 5-fold cross-validations of the model that includes only economic and demographic features. The mean of these accuracies is 0.711, and the standard deviation of the accuracies is 0.084. The mean accuracy is 7.1% lower than that of the first model that included all variables, and the standard deviation is also higher. This indicates that this model has a lower accuracy and a higher variance than Model 1; 5-fold cross-validation indicates that this model does not perform as well as the model with all features.

Fold	Accuracy
1	0.824
2	0.765
3	0.647
4	0.588
5	0.733

Table 5.3: 5-fold Cross-validation Accuracies for Model 2

The 5-fold cross-validation gives an indication that the model is expected to classify new farmers markets correctly 71.1% of the time.

The 5-fold cross-validation indication is that this model is not as good a classifier as model 1. This will aid in model analysis as we additionally look at the performance indications from the confusion matrix. Table 6.4 contains the confusion matrix on the test data that has been set aside to further evaluate model performance.

	Predicted Market Failure	Predicted Market Success
Actual Market Failure	18	3
Actual Market Success	5	10

Table 5.4: Confusion Matrix for Model 2

The confusion matrix in Table 5.4 shows that there are 18 true negatives, 3 false positives, 5 false negatives and 10 true positives. The accuracy of the model on the test set is calculated to be

$$\frac{18+10}{18+5+3+10} = \frac{28}{36} = 0.778.$$

According to the confusion matrix, this model does a fairly good job of classifying farmers market success at 77.8% accuracy. It is only 2.8% less accurate than Model 1's confusion matrix. While it is not inappropriate to say that the confusion matrix tells us that the model is expected to correctly classify a new farmers market 77.8% of the time, this illustrates the importance of doing 5-fold cross-validation in addition to setting aside a test set. Based on 5-fold cross-validation, model 2 is actually much less accurate than Model 1 than what the confusion matrix indicates.

The lower accuracy and higher variance of this model may indicate that farmers market specific variables play an important role in classifying markets' success since excluding these variables had a negative effect on accuracies. This will be further addressed in Model 3.

5.2.3 Model 3 (Best Model)

Again, the data was split into 70% training data and 30% test data. Note that the training data and test data are still made up of the same observations as in the previous models.

Using the ANOVA F-Value method to determine if individual features have a relationship with market success, four features with F-Values greater than five were kept in the model. The features kept were number of vendors, population size, number of years the market has been established, and median age. Using the grid search function on these data to optimize the mean accuracies of the 5-fold cross-validation results in-hyper parameters of kernel=Linear and C=100.

A linear kernel indicates that the data is linearly separable. A linear kernel results in a simpler model and allows for more interpretation of the features' effects on the classifier's decision of a successful market than did the two previous models with rbf kernels.

The 5-fold cross-validation accuracies for Model 3 is displayed in Table 5.5. The mean of these accuracies is 0.795, and the standard deviation of the accuracies is 0.026. The mean accuracy is 1.3% higher than that of the first model, which included all variables, and the standard deviation is lower. This indicates that the model has a higher accuracy and a lower variance than Model 1. The 5-fold cross-validation results indicate that Model 3 performs slightly better than Model 1, which included all features.

Fold	Accuracy
1	0.824
2	0.824
3	0.765
4	0.765
5	0.800

Table 5.5: 5-fold Cross-validation Accuracies for Model 3

Based on the 5-fold cross-validation, the model is expected to classify new farmers markets correctly 79.5% of the time.

Testing the model on the test data, which the model has not yet seen, results in the confusion matrix in Table 5.6.

	Predicted Market Failure	Predicted Market Success
Actual Market Failure	18	3
Actual Market Success	3	12

Table 5.6: Confusion Matrix for Model 3

The confusion matrix in Table 5.6 shows that there are 18 true negatives, 3 false positives, 3 false negatives and 12 true positives. The accuracy of the model on the test set is calculated as

$$\frac{18+12}{18+3+3+12} = \frac{30}{36} = 0.833.$$

This is the highest accuracy of the confusion matrix among all three models. Solely looking at the confusion matrix results, it is expected that the model will correctly classify 83.3% of new farmers markets' success correctly. Note here that because the data set is relatively small, having one more or one less misclassification can substantially change the accuracy of the model on the test set. Having a small data set is another reason that it is important to rely on both 5-fold cross-validation and the confusion matrix from the test set.

Model 3, which only includes four features (number of vendors, population size, number of years the market has been established, and median age), performs better than the model with all features. Population size and median age are the only remaining features that were demographic in nature. This provides some evidence that economic and demographic predictors may not be the best predictors for farmers market success since removing a majority of the economic and demographic indicators improved the model.

Figure 5.4 displays the weight that each feature that is left in the model has on the classification decision. The magnitudes of the weights show that number of vendors had the largest effect on the classification decision by a large margin. This is followed by median age, population size and number of years the market has been established.

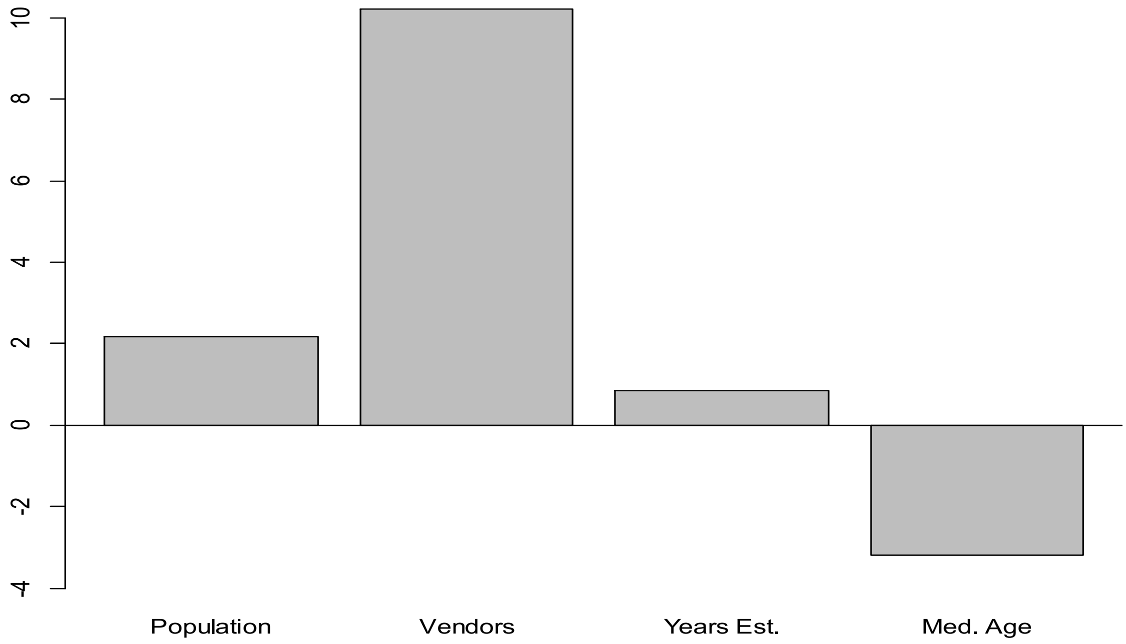

Figure 5.4: Weights of Features on Classification Decision

CHAPTER SIX: DISCUSSION

6.1 SVM vs. Logistic Regression

The results were robust across both methods. That is, both methods resulted in very similar indications regarding farmers market success. The only difference in the final model under each method is the inclusion of median age in the SVM model. Here is a good place to note that SVM and logistic regression have similar loss functions; because of this, it would be expected that the models would perform similarly, and the empirical results from this study support this theoretical implication.

Along with analysis of predictors of farmers markets success, this research is seeking to analyze machine learning methods versus more traditional methods. Even with similar performance across both methods, a few things are apparent. Logistic regression offers much more interpretable results than SVM, especially when a kernel is being used with the SVM. Since the goal in most traditional statistical and economic studies is to have interpretability of the estimates, when models have similar results, it makes sense for statisticians and economists to continue to use traditional logistic regression models. If the goal of research is to have a model that is focused solely on predictability power, SVMs have a slight advantage over logistic regression. Logistic regression performs best when data is linearly separable, but SVM can create good models that do not have linearly separable data by using kernels. For example, Model 1, which included all variables, was found to be poorly fitting according to the Hosmer and Lemeshow Test for the logistic regression model, but with the rbf kernel using the SVM, the model still showed to be a good predictor for farmers market success. Logistic regression is an important tool, and this research concludes that it should still be the primary tool for economists when interpretability of variables is desired. What the SVM model, as well as other classification models, offers is an alternative for when logistic regression does not perform well or in cases when predictability is more important. However, traditional use of logistic regression does not include the separation of training and test data. This is one area of improvement the traditional use could take from machine learning. Assuming enough data is available, it is good practice to separate the data. It is an easy empirical way to see how well the model will perform on data the model has not yet seen. Economists want their models to be appropriate for data their model has not already seen. They do not want to make an over-fitted model, but this risk is taken when data is not separated into training and test sets. Splitting data leads to the same interpretability of variables but with more confidence that the interpretations will hold on unseen data.

6.2 Farmers Markets

The results from both logistic and SVM modes indicate that market-specific variables (number of years markets have been established and number of vendors) are better predictors of farmers market success than the economic and most demographic variables. It makes sense that market-specific variables may be more important than other variables. Obviously the longer a market has been open, the more likely it is to be successful since poor performance of a market over several years would likely result in its shutting down. The number of vendors also makes sense as a useful predictor. Consumers tend to like variety, and having more vendors offers that variety; it also makes sense that a market that is doing well will attract more vendors. Number of vendors is a good predictor of market success, but more research would need to be conducted to uncover a causal relationship because the simultaneity would need to be unwrapped. Despite concerns that might arise because of endogeneity in the model, when policy makers are handing out grants, number of vendors, as well as other variables, are predetermined since existing farmers markets are already in place. Population size is the other variable that appeared across both methods' best models. This also makes sense because if markets have a greater number of people in close vicinity to them, it makes sense that they would perform better than those with fewer potential customers nearby. Additionally, both models showed that population size offered more weight in the decision-making process than the number of years the market has been established. Median age appeared as an important predictor for the SVM model but did not remain in the logistic regression model.

Perhaps more interesting than the variables found to be important were the variables that were found to be unnecessary. Model 2, which included only economic and demographic variables, performed poorly under both SVM and logistic regression in comparison with the best model of each method (Model 3). Besides population size, and median age in the SVM model, the economic and demographic variables offer little new information when market-specific variables are included in the model. At the beginning of this exploratory study, it was expected that economic and demographic variables would be important predictors, but this was found not to be the case, at least at the zip-code level. This research does not completely rule out the ability of economic and demographic variables to be useful predictors in general. The minimal impact they have is only represented at the zip-code level. It is possible that in order for these variables to offer more information, they need to be in closer vicinity to the market. Instead of zip code level data, perhaps data that came from within a ten-mile radius of the market would be more informative.

Nevertheless, this research does begin to question whether economic well-being is important in predicting farmers market performance. Perhaps the Kentucky Double Dollars program, which doubles the amount of money SNAP and EBT recipients plan on spending at the market, makes farmers markets accessible across all income levels. Thus, the economic condition near the market could be irrelevant. Again, this is only a theory that future research could explore further.

To improve the predictability power of the models, more market-specific data could be useful. If economic and demographic variables are not the best predictors, as this research suggests, it may be advantageous to try new models that incorporate variables such as whether or not the market has a parking lot, whether or not they advertise, the distance to the nearest grocery store, whether or not it is covered with a pavilion, etc. Future research could also examine panel data in order to take survivorship into account. This research inherently suffers from survivorship bias since all markets used in the study were operating in 2017. Having panel data would allow research to account for this survivorship bias by making it possible to include markets that ultimately failed.

This research gives support to the idea that farmers markets are not destined for success or failure based on the economic conditions near which they find themselves. Future research could use similar methods that included more market-specific variables. If it is found that a market’s success is better predicted by market-specific traits, it would offer support to the notion that markets have the ability to control their own fate. Future research could examine how farmers markets have performed during recessions. If farmers markets experienced little impact from economic recessions, it would provide further evidence that economic conditions do not play a substantial role in farmers market success.

In summary, this exploratory study adds to the limited quantitative research that is currently being done with farmers markets. Furthermore, it is the first study that uses both traditional statistical methods as well as machine learning methods to analyze farmers markets' success. It gives evidence that both models have similar results, but that logistic regression should be used when interpretability is important, while SVM can perform better if predictability is the focus and data is not linearly separable. Finally, it provides a framework for future research to analyze other variables' ability to predict farmers market success. It is the hope that this research acts as a starting point for research that looks at the impacts of market-specific economic and demographic traits in predicting farmers market success. This research built a model that demonstrated 83.33% accuracy; it is hoped that future research will use this framework to create a model that is an even better predictor of success. This research offers a tool that can help in policy decision by government and the foundation of a tool for farmers market managers to increase the likelihood of farmers markets' success.

REFERENCES

Agresti, Alan. (2007). *An Introduction to Categorical Data Analysis: Second Edition.* John Wiley and Sons, Inc.

Bonanno, A., Berning, J., & Etemaadnia, H. (2017). Farmers Market Locations and Their Determinants: An Empirical Analysis in New England. *Agricultural and Resource Economics Review, 46*(3), 479-506. doi:10.1017/age.2016.43

Brown, A. (2002). Farmers' market research 1940±2000: An inventory and review. *American Journal of Alternative Agriculture,* 27(4), 167-165. doi: 10.1079/AJAA200218

Cristianini, N., & Schölkopf, B. (2002). Support vector machines and kernel methods the new generation of learning machines. *AI Magazine*, *23*(3), 31–41.

Cummings, H., Kora, G., & Murray, D. (1998). Farmers' markets in Ontario and their economic impact. Retrieved from http://hcaconsulting.ca/wp-content/uploads/2017/03/1999-Farmers-Markets-Ontario-Economic-Impact.pdf

Gumirakiza, J.D., Curtis, K.R., & Bosworth, R. (2014). Who attends farmers' markets and why? understanding consumers and their motivations. *International Food and Agribusiness Management Review,* 17(2): 65–82. Retrieved from https://www.ifama.org/resources/Documents/v17i2/Gumirakiza-Curtis-Bosworth.pdf

Henneberry, S.R., Augustini, H. N., Taylor, M., Mutondo, J.E., Whitacre, B., & Roberts B.W. (2008). The economic impacts of direct produce marketing: a case study of

Oklahoma's famers' markets. *Southern Agriculture Economics Association Annual Meeting Paper Presentation.* doi: 10.22004/ag.econ.6785

Hughes, D. W., Brown, C., Miller, S., & McConnell, T. (2008). Evaluating the economic impact of farmers' markets using an opportunity cost framework. *Journal of Agricultural and Applied Economics*, 40(1), 253–265. doi: 10.1017/S1074070800028091

Hughes, D.W., & Isengildina-Mass, O. (2015). The economic impact of farmers' markets and a state level locally grown campaign. *Food Policy,* 54, 78-84. doi: 10.1016/j.foodpol.2015.05.001

Otto, D., & Varner, T. (2005). Consumers, vendors, and the economic importance of Iowa farmers' markets: an economic impact survey analysis. *Leopold Center Pubs and Papers,* 145. Retrieved from http://lib.dr.iastate.edu/leopold_pubspapers/145

Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *JMLR*. 12(85):2825−2830, 2011.

Pedregosa *et al.,*. (n.d.). Scikit-learn User Guide. Retrieved April 7, 2020, Retrieved from https://scikit-learn.org/stable/user_guide.html

Schmitz, E. (2010). *Farmers' Markets in Kentucky: A Geospatial, Statistical, and Cultural Analysis* (Masters Theses & Specialist Projects Paper 214). Retrieved from http://digitalcommons.wku.edu/theses/214

Singleton, C. R., Sen, B., & Affuso, O. (2015). Disparities in the availability of farmers markets in the United States. *Environmental Justice,* 8(4): 135–143.

Wolf, M.M., Spittler, A., & Ahern, J. (2005). A profile of farmers' market consumers and the perceived advantages of produce sold at farmers’ markets. *Journal of Food Distribution Research,* 36(1): 192–201. doi: 10.22004/ag.econ.26768

Yosick, B. (2008). Economic Impact of Portland’s Farmers Markets. Retrieved from https://www.portlandoregon.gov/bps/article/236588

Zhang, G. (2018). What is the kernel trick? Why is it important? [Blog post]. Retrieved from https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-8a98db0961d

APPENDIX

As referenced in the footnote on page 9, Table A displays estimated coefficients for the logistic regression model that uses the median of profit per vender, as opposed to gross profit, as the cutoff for binary classification of the dependent variable (success or failure). Chi-square test statistics are included in the parenthetical and *, **, *** represent significance at the 0.1, 0.05 and 0.01 level, respectively. An interesting finding was that Model 3 no longer includes number of vendors, but does include the percentage of people with at least a bachelor's degree. The model is still robust to a changing definition of success in finding that number of years the market has been established and population are still included in the model.

	Model 1	Model 2	Model 3
Population	.000058	.000065	0.000066
	(4.4567)**	(6.0694)**	(8.3409)***
Unemployment (%)	-0.0480	-0.0734	--
	(0.3366)	(0.8269)	
Bachelor's (%)	0.0532	0.0558	0.0544
	(2.8625)*	(3.5246)*	(5.3795)**
White (%)	-0.0119	-0.00821	--
	(0.1237)	(0.0579)	
HMI ($)	-0.00002	-0.00002	--
	(0.3874)	(0.9177)	
SNAP (%)	-0.00999	-0.0118	--
	(0.0790)	(0.1145)	
Median Age	-0.0693	-0.0759	--
	(0.7691)	(0.9355)	
Vendors	0.00857	--	--
	(0.1937)		
Years Established	0.0249	--	0.0296
	(1.8258)		(2.8110)*

Table A: Estimated Coefficients with Profit per Vendor used for Classification of Market Success