# Springboard Capstone Project

# Problem:

Insurance companies rely on policy revenue for their profits. When customers cancel policies prior to a year, there is lost revenue. It is difficult to prevent this loss of revenue because we do not know who will cancel ahead of time.

# Uses:

An insurance company could potentially use this information for two purposes. First, they could use it to identify the leads that are likely to cancel and do not buy those leads. That money would be better spent of buying leads for customers who will be less likely to cancel.

Second, after a policy is purchased, identify those who are most likely to cancel. Then have targeted interactions with them aimed at preventing cancellations e.g. payment reminders, phone calls to ensure satisfaction.

# Data:

Data is available from a small insurance company. The data being used contains variables that are relevant to each lead in the data set: Demographic characteristics (gender, age, marital status, credit, geography), previous insurance information (insured vs. not insured, policy type, bi limit), car (make, annual mileage, number of vehicles), policy purchase information (same day vs. follow up sale)

# Approach:

The data analysis took place in three phases, which include data wrangling, exploratory analysis, and predictive modeling. In the data wrangling phase, the data from three different sources was combined. The lead data information included all of the information we know when a lead is purchased, the purchase data included all of the purchase information, and the cancel data gave us our target variable.

Another aspect of data wrangling was be examining missing data and determining the best course of action. Rows of the data set that contained multiple missing values were deleted from the data set. In other cases, the missing data indicated that there was no cancel or the customer didn’t want something revealed. In those cases, there was an other/missing category put into categorical variables. In cases of a numeric variable, the mean was using to fill the missing values.

* \*Date Transformation: Create categories for examining date: Month, Day of week
* \*Any other issues that need to be addressed before analysis?

1. Exploratory Analysis:

* \*Which variables correlate with cancels?
* \*Graphs to show variables of interest

1. Predictive Analysis:

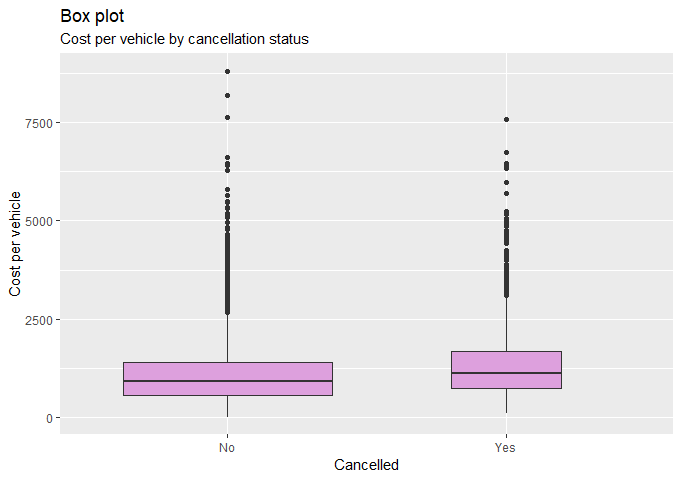
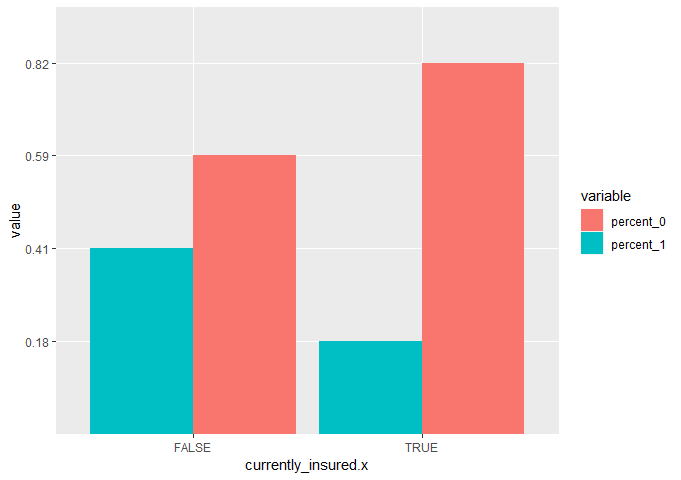
* \*Anaysis to show which variables predict cancels
* \*Analysis to show which variables predict cancels by lag time
* \*Anaysis to show which variables predict cancels by cancel reason

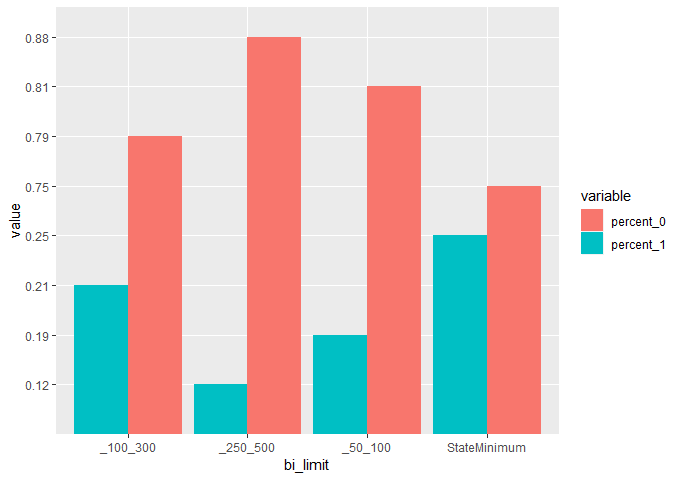
# Results:

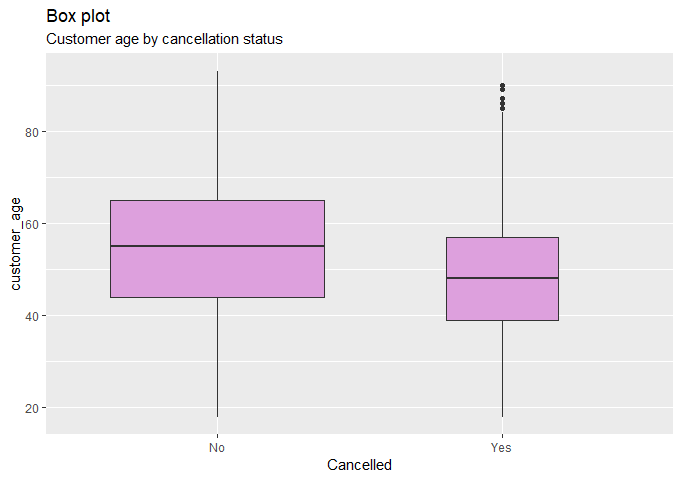
Chi-squared test for independence was done for all categorical independent variables, while t-tests were done for all numeric independent variables. The following indicates the statistical significance levels of all the variables.

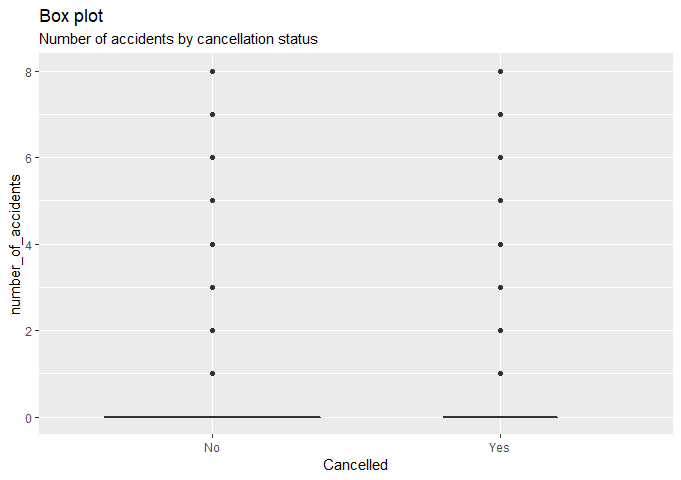
|  |  |
| --- | --- |
| **Variable Name** | **Statistical Significance** |
| Agent | 2.2E-16 |
| bi\_limit | 2.2E-16 |
| cost\_per\_vehicle | 2.2E-16 |
| currently\_insured.x | 2.2E-16 |
| customer\_age | 2.2E-16 |
| Education | 3.406E-12 |
| Gender | 2.088E-15 |
| lead\_seller | 2.2E-16 |
| LQ\_FB\_PREM\_AMT | 2.2E-16 |
| marital\_status | 1.373E-14 |
| Months\_Coverage | 2.2E-16 |
| no\_of\_vehicles | 0.00009354 |
| number\_of\_accidents | 0.02133 |
| number\_of\_claims | 0.1504 |
| previous\_policy\_type | 2.2E-16 |
| risk\_profile | 2.2E-16 |
| SameDay\_vs\_Followup | 2.025E-13 |
| State | 2.2E-16 |
| vehicle\_ownership | 0.3385 |

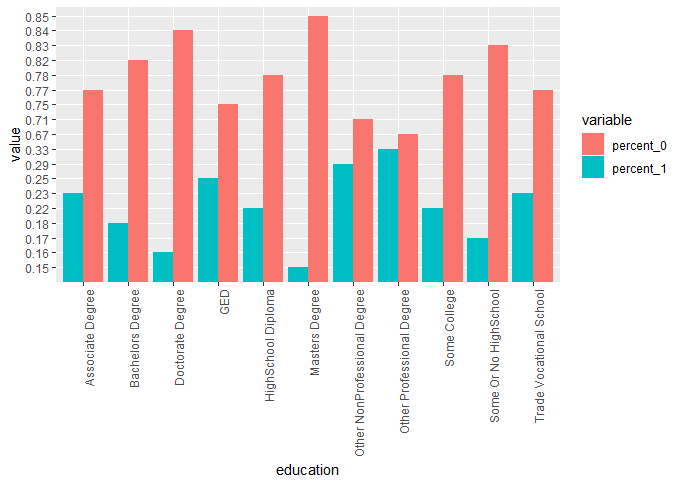
The following are graphs showing the relationships between the independent and dependent variables:

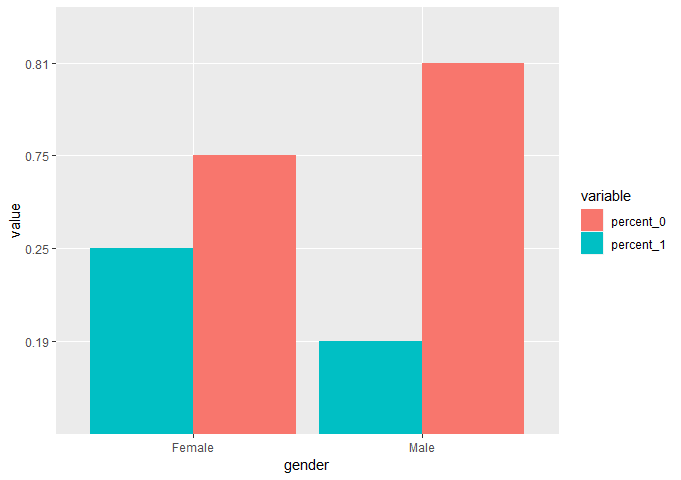


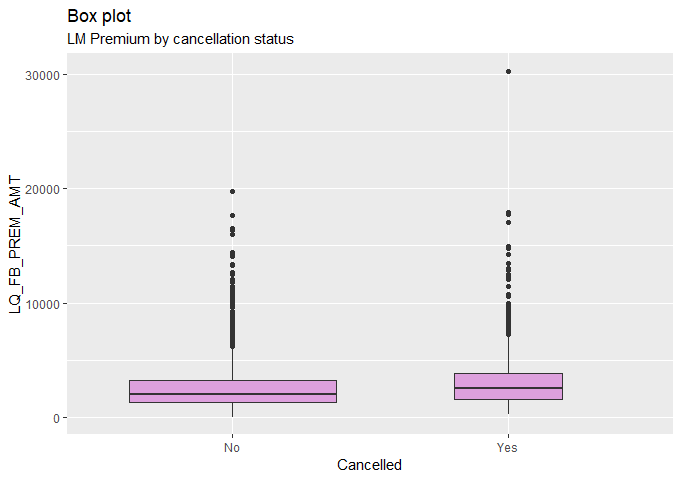


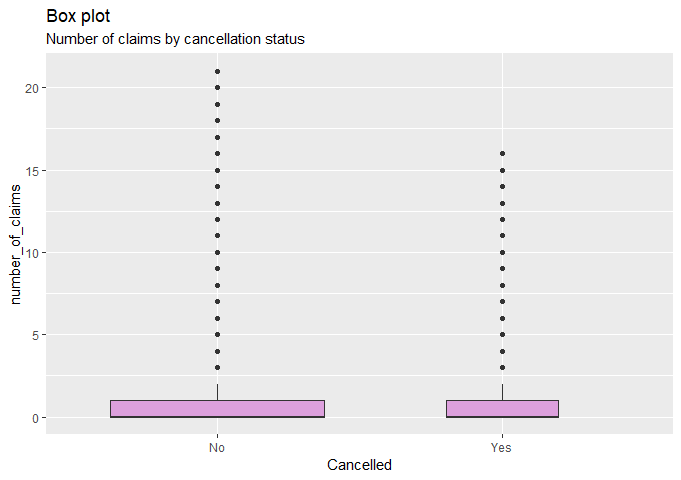


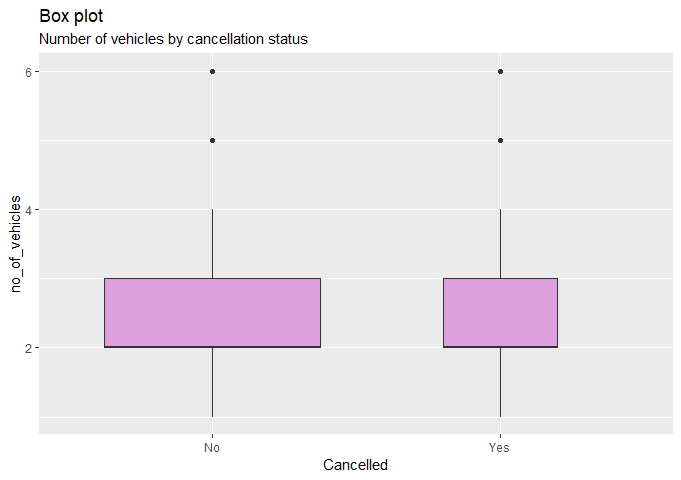


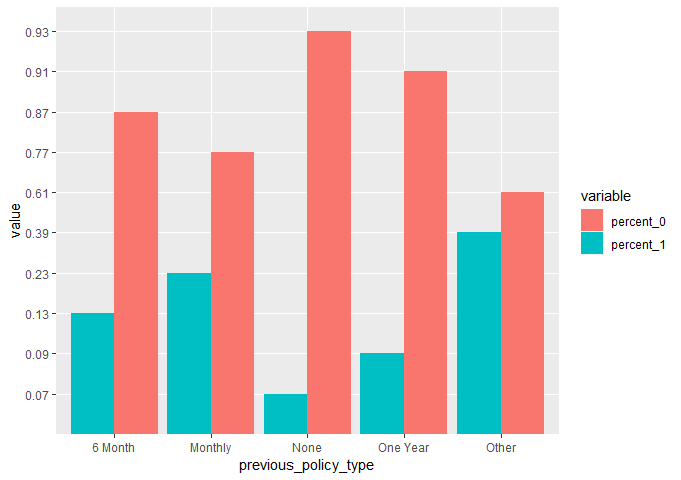


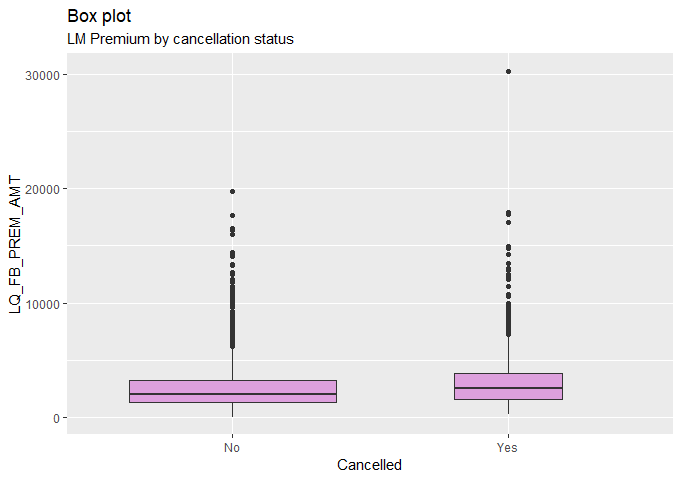




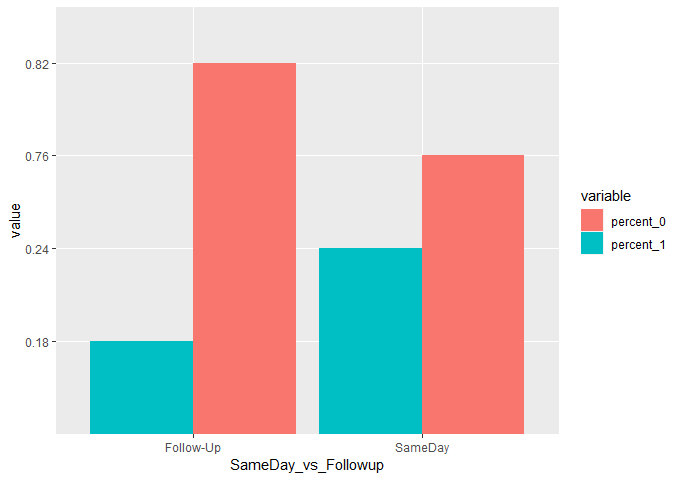


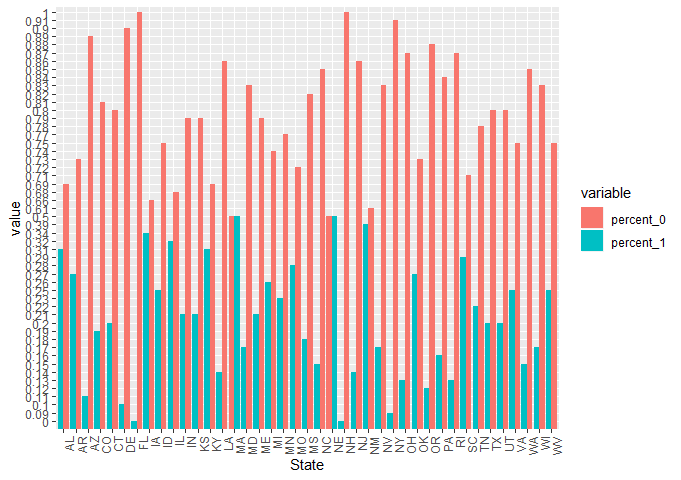


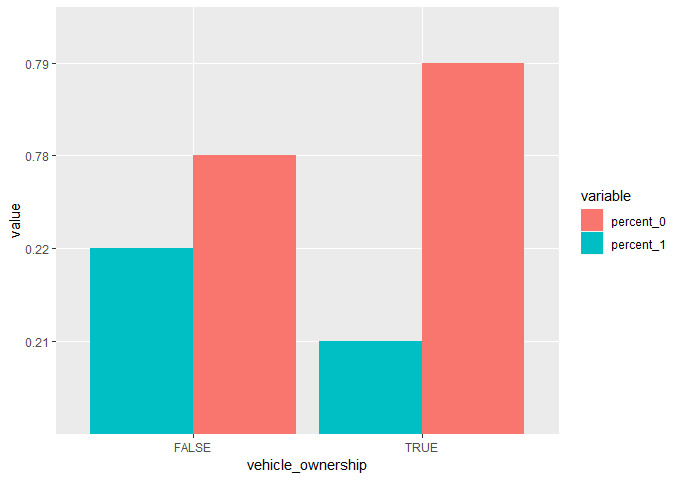


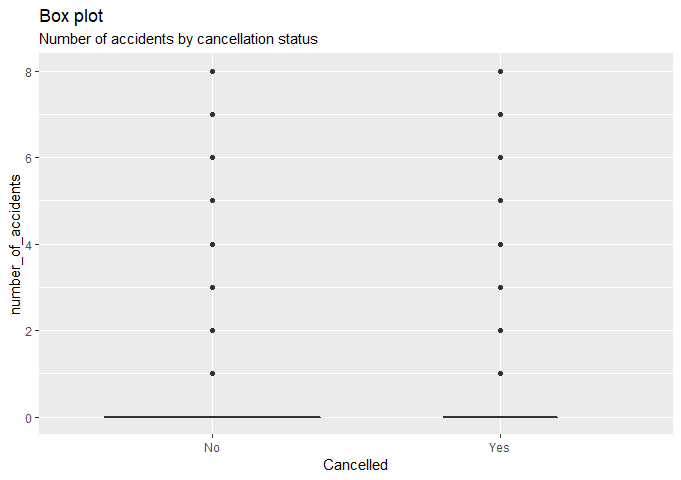


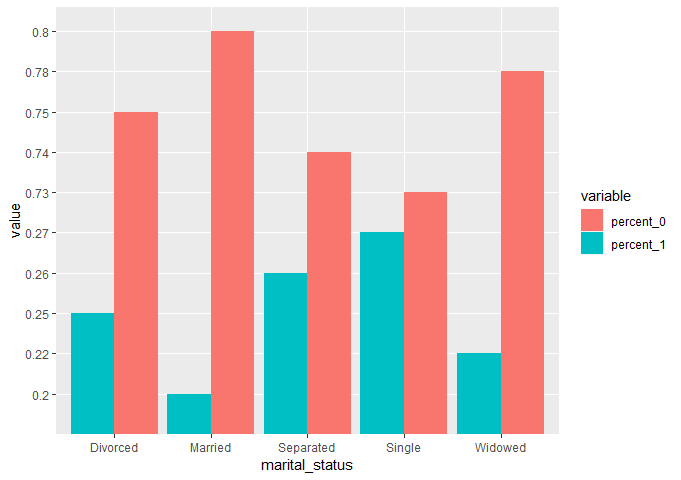












# Modeling:

A stepwise regression was done to define which variables should be included in the models. This found that the model would be best when the following variables are included:

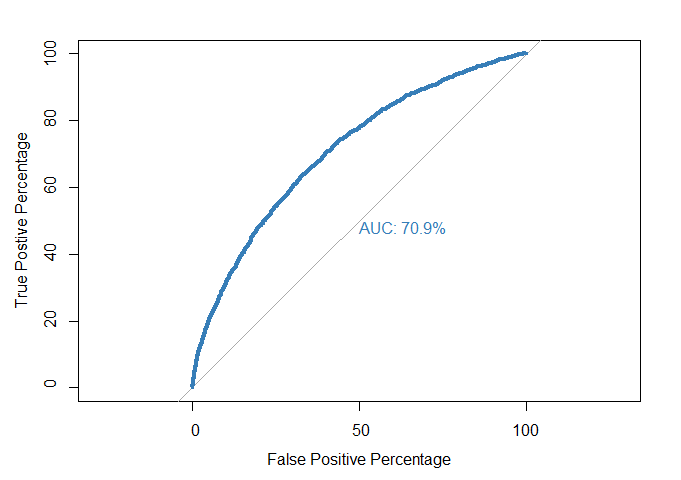
Cost per vehicle, marital status, bi limit, currently insured or not, customer age, gender, previous policy type, same day vs. follow-up sale, state, and lead seller.

State and lead seller both caused errors in the models due some levels of each variables having low volume. They were removed from the models.

Three types of models were run: A general linear model, random forest, and svm. The following indicates the results of each type of model when trained on 75% of the data and tested on the remaining 25%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | GLM | Random Forest | SVM |
| Accuracy | 0.7857 | 0.7839 | 0.7831 |
| P-Value | <2e-16 | <2e-16 | <2e-16 |
| Confidence Interval (95%) | 0.7698, 0.8011 | 0.7679, 0.7993 | 0.7671, 0.7985 |
| Sensitivity | 0.9887 | 1.000000 | 1.0000 |
| Specificity | 0.0529 | 0.003413 | 0.0000 |

The general linear model has the best accuracy level. An ROC curve and AUC metric was calculated or the general linear model to confirm it was the best model to use. Below is the ROC curve with the AUC metric.



These results are positive and confirm the use of the general linear model.

# Conclusions:

The results show that the general linear model is a good fit for the data. These results will be used to predict policy cancellations. After additional data is collected, it would be useful to reassess the model to see if it needs to be adjusted.