**Joseph Deal**

**Student ID: 000655053**

**Western Governors University**

**Logistic Regression**

**For Predictive Modeling**

**Part I:**

**RESEARCH QUESTION**

**A.**

**1.** Utilize the cable company data set to determine which customers are likely to terminate service?

**2.** The goal is to define which independent variables drive the decision to end their cable service. Providing this information would be critical in reducing 'churn' or customer turnover. Predicting which customers are most likely not to renew their service and would provide the means to provide an incentive for the customer to stay.

**Part II:**

**METHOD JUSTIFICATION**

**B.**

**1.** Logistic regression analysis is depended on a few assumptions regarding the dataset. The most obvious assumption is that the response or dependent variable is binary. In other words, the variable you want to predict only has two values.A yes or no, zero or one, pass or fail, male or female, hot or cold are examples of binary results.

Another assumption is that all the observations in the dataset are independent. Each record in the dataset is unique, not a repetition of another record or a record dependent on another record to be complete or related in any way.

A third assumption is multicollinearity. Multicollinearity is the correlation between variables. This requirement is that none of the independent variables should be highly correlated. The multicollinearity of multiple variables in the dataset would sway the analysis towards a specific result. This result would not be providing independence in the variables.

A fourth assumption is that the dataset has enough records or observations. The number of observations should be large enough to evaluate and come to a conclusion. If the dataset were too small, the conclusion would be biased because of the lack of data.

Another assumption is that there are no extreme observations. These extreme observations would need to be moderated or removed due to the large impact on the regression analysis. This extreme outlier impact is due to the utilization of evaluating residuals during the linear regression process.

**2.** Using python as the main tool for analysis provides significant opportunities for utilization, sharing, and applications. Many development environments can utilize Python from the most robust, to an online text editor. Jupyter notebooks are an invaluable tool in which python programming can be used and shared readily. Google Colabs (colabs.research.google.com) is another environment that can be accessed from anywhere. Almost any device can access Colabs if it has internet service.

Python has numerous tools that can be added to its environment to improve performance and add algorithms to enhance data analysis. Some of those tools include Numpy, Pandas, Scikit-Learn, Seaborn, Matplotlib, Statsmodel, Pyspark, Spark, and many more.

**3.** Logistic regression analysis is very appropriate with a dependent variable that has binary characteristics. In this case, whether a customer will churn or not is a yes or no result. The churn variable provides an excellent binary product for logistic regression analysis.

**Part III:**

**DATA PREPARATION**

**C.**

**1.** Data Preparation will consist of several steps. Focusing on values that programming will throw errors includes missing values like, Nulls, NaNs, and any strangely formatted data. There is a multitude of ways to deal with these types of issues. Replacing Nulls and NaNs for a numeric field can be as simple as replacing with zeros or as complex as calculating the values' average based on specific other characteristics of the dataset.

Checking for data types found in each feature is critical in analysis performance as well. Verifying data types will prevent another set of issues that programming and analysis will not interpret correctly. For instance, a feature may appear to be a currency field. The field includes the dollar sign, and the data type for the feature may be text instead of a number.

Another data prep issue is determining if the numeric field is continuous or discrete or if the character fields are categorical or free text. The understanding of the structure of the field helps in exploring extremes and rectifying any issues found.

If there is a critical record identifying missing information, the record may be deleted if there are many records in the dataset.

Duplicate records need deletion. There were no duplicate records found.

Checking for extremes in any of the column types is an excellent way to evaluate continuity. Many extreme entries are errors and need to be addressed. These extreme values could significantly impact the data analysis, particularly true if the dataset is small. Handling these extreme values is typically a logic check. If the data clearly makes no sense or is impossible, then a decision on how to handle the issues come to the forefront. Do you delete the record? Do you average out the value if it is a number? Do you make it zero or the mode value? There are lots of options.

For analysis purposes, text-based classification variables will need to be enumerated. Pandas.get\_dummies method to encode the text classification to numbers will be used on the following fields, PaymentMethod, Contract, Marital, InternetService, Port\_modem, Phone, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies.

**2.** A quick look through the table finds that numerous fields will not be part of the general analysis. These include Lat, Lng, UID, Interaction, Customer\_id, CaseOrder, TimeZone, Job, Zip, County, State, City, and Area.

A quick thought on the locations and their relevance is enveloped in the zip field. The zip encompasses all the state, city, area, county, lat and lng, and TimeZone. Looking at the zip field distribution in the below chart shows an even distribution of customers from the 00xxx's to the 88xxx's. This reflects that no one area is suffering from churn anymore than any other. The slight dip in the 00xxx's and the 88xxxs ranges reflect the lack of population in the northern New England area and the Rocky Mountain states.

|  |
| --- |
|  |
| Chart 1 – Zip Code Distribution |

There are 10,000 records for every field. There are no records that have blanks, NaNs, or Null values.

Looking for extreme values are easily seen in the following table showing field name, min, max, average, and quartiles.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Income** | **Tenure** | **MonthlyCharge** | **GB/Year** | **Contacts** | **EquipFail** | **Items** |
| **Mean** | 53.07 | 39806.92 | 34.53 | 172.62 | 3392.34 | 0.99 | 0.39 | 3.5 |
| **std** | 20.69 | 28199.91 | 26.44 | 42.94 | 2185.29 | 0.98 | 0.63 | 1.0 |
| **min** | 18 | 348.67 | 1.00 | 79.97 | 155.5 | 0.00 | 0.00 | 1.00 |
| **25%** | 35 | 19224.71 | 7.92 | 139.97 | 1236.47 | 0.00 | 0.00 | 3.00 |
| **50%** | 53 | 33170.61 | 35.43 | 167.48 | 3279.53 | 1.00 | 0.00 | 4.00 |
| **75%** | 71 | 53246.17 | 61.48 | 200.73 | 5586.14 | 2.00 | 1.00 | 4.00 |
| **max** | 89 | 258900.70 | 71.99 | 290.16 | 7158.98 | 7.00 | 6.00 | 7.00 |

In the table, the Age range goes from 18 to 89, which seems very reasonable. 75% of the users are under the age of 71. Income ranges from $348 to $258,900. The low end seems odd, but it could be college-age customers who may not earn a great deal and are subsidized by parents.

The Tenure field seems reasonable as well, with extremes of 1 to 72 months. Monthly charges from $79.97 to $290.16 appear to be a reasonable spread. Contacts align nicely with equipment fails (EquipFail) from min to max with an additional contact than an equipment failure.

The target field is churn. This field reflects the number of customers of all the total customers that have terminated their service. It appears that approximately 26.5% have terminated their service in the last year, as noted in Chart 2.

|  |
| --- |
|  |
| Chart 2 – Churn Percentages |

It was determined that Martial, Gender, InternetService, PaymentMethod, Item1 through Item8, OnlineBackup, OnlineSecurity, Multiple, Phone, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, paperlessBilling, TimeZone, Area, County, Zip, State, City, Churn are all categorical fields. Most are binary such as Churn, OnlineBackup, OnlineSecurity, Multiple, Phone, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, paperlessBilling. Simultaneously, the balance of categorical fields has multiple nominal values except for the Item1 thru Item8 fields, which are ordinal values from the survey given after each customer interaction.

The customer interaction survey results appear to be very consistent. See the charts for items1 through item8 below.

|  |
| --- |
|  |
| Chart 3 – Survey Question Distributions |

All the categorical variables have a consistent and reasonable selection of values. Marital status has 'Widowed', 'Married', 'Separated', 'Never Married', 'Divorced'. PaymentMethods variable has 'Electronic Check', 'Mailed Check', 'Bank Transfer(automatic)', 'Credit Card(automatic)' and standard values. Gender has 'Male', 'Female', 'Nonbinary'. InternetService has 'Fiber Optic', 'DSL', and 'None'. Contracts have three values of 'Month-to-month', 'Two Year', 'One Year'.

**3.** After loading the data into a pandas dataframe in python, the steps used to prepare the data for analysis include the following steps:

* A quick review of the data to evaluate what the data looks like using the following code:
  + df.info – which gives a look at the top 5 rows and the bottom 5 rows of the dataset loaded into the dataframe df.
* A quick statistical overview of the dataframe is accomplished using:
  + Df.describe()
    - This provides a count, mean, standard deviation, minimum, 1st quartile, 2nd quartile, 3rd quartile and max values for all the numerical variables in the dataframe.
    - This is where extreme values should be visible and addressed.
* Review all the column names.
  + Df.columns will present the list of column names in the dataframe.
* Review all the variable data types.
  + Df.dtypes will return a list of all the variable and the data types associated with each, such as int64, object, float64.
  + A quick review here will identify the numeric fields from the text fields.
* Review general contents of all variables.
  + Using df.info() will count each variable, including the variable name, the number of Non-Null values, and the data type of the field.
  + Info() is a quick and easy way to find missing values and where they are in the dataframe.
* Review for missing information
  + Using df.isna().any() will review each variable and check for NaNs, Nulls, NaTs for each variable. It returns a list of variables with a True if there are missing values or False if there are no missing values.
* Review all the object variables and values
  + Df[‘Marital’].unique() will return all the unique values in the variable.
  + Reviewing those unique values will spot any misspellings or extreme values for the variables.
  + Used .unique() on all categorical variables such as Marital, PaymentMethod, InternetService, Contract, Gender, TechSupport, DeviceProtection, OnlineBackup, OnlineSecurity, Multiple, Port\_modem, Tablet, Phone, PaperlessBilling, StreamingTV, StreamingMovies, and Techie.
  + There were no extreme or irrational values in these variables.

**4.** Univariate visualizations:

|  |
| --- |
|  |
| Chart 4 – Churn Percentages |

|  |
| --- |
|  |
| Chart 5 – Payment Method Distribution |
|  |

|  |
| --- |
|  |
| Chart 6 – Internet Service Distribution |
|  |

Bivariate visualizations:

|  |
| --- |
|  |
| Chart 7 – Churn by Marital Status |
|  |
|  |
| Chart 8 – Churn by Gender |
|  |
| Chart 9 – Churn by Internet Service Type |

**5.** Copy of the cleaned dataset is in the file named prepared\_churn\_data.csv.

**Part IV:**

**MODEL COMPARISON AND ANALYSIS**

**D.**

**LOGISTIC REGRESSION WITH ALL PREDICTORS**

1. The following code provided the structure and output for the first model.

(Comments are bold. Code is plain text.)

**# Loaded the proper modules from sklearn**

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import classification\_report, confusion\_matrix

**# Split the dataset into a training and testing set. Using an 80/20 testing/training split.**

X\_train, X\_test, y\_train, y\_Test = train\_test\_split(X\_features, y\_features, test\_size=0.80, random\_state=234)

logmodel = LogisticRegression(solver='liblinear', random\_state=0)

logmode.fit(X\_train, y\_train)

**# Run the predict method with the test data to evaluate the logistic model with all the predictors.**

Logmodel.predict(X\_test)

**# Run the score method to calculate the precision of the logistic model.**

Logmodel.score(X\_test,y\_test)

**# Results of the scoring was 0.844. An 84.4% prediction rate for churn.**

**# For more details we will display the confusion matrix and calculate precision, accuracy and recall values.**

confusionmatrix = confusion\_matrix(y\_Test, logmodel.predict(X\_test))

fig, ax = plt.subplots(figsize=(8, 8))

ax.imshow(confusionmatrix)

ax.grid(False)

ax.xaxis.set(ticks=(0, 1), ticklabels=('Predicted 0s', 'Predicted 1s'))

ax.yaxis.set(ticks=(0, 1), ticklabels=('Actual 0s', 'Actual 1s'))

ax.set\_ylim(1.5, -0.5)

for i in range(2):

for j in range(2):

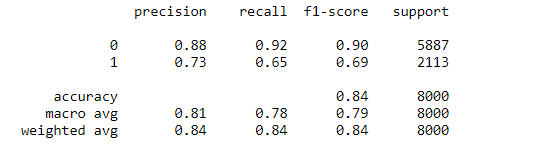
ax.text(j, i, confusionmatrix [i, j], ha='center', va='center', color='red')

plt.show()

**# Produces the following Chart:**

|  |
| --- |
|  |
| Chart 10 – Confusion Matrix ALL PREDICTORS |

Print(classification\_report(y\_test, logmodel.predict(X\_test)



# Looking at the classification report, it is notable that the 0 prediction is much higher than the 1. With the 1 being the actual churning customers, the overall prediction rate is 69%.

1. There were 42 features evaluated during the logistic regression. Not all those predictors were helpful. Determining which features to focus on and which to ignore will require more information. Utilizing a correlation heatmap is useful, but it is a little difficult to read due to the number of features. Please refer to the chart below.

|  |
| --- |
|  |
| Chart 11 – Correlation Heat Map |

The important part of the heatmap is the Churn\_Yes and Churn\_No rows highlighted by a yellow box. Along the yellow box there are 'warm' areas noted by the green arrows. These areas coincide with features Tenure, MonthlyCharge, Bandwicth\_GB\_Year, Contract\_MonthtoMonth, StreamingTV and StreamingMovies.

Another python statistic module called statsmodel has regression methods that have a robust statistical analysis. To verify the results using the scikitLearn LogisticRegression model, I installed statsmodels.api and ran statsmodels Logit method on the same dataset. The summary results are noted below.

|  |
| --- |
| Logit Regression Results  ==============================================================================  Dep. Variable: Churn\_Yes No. Observations: 2000  Model: Logit Df Residuals: 1958  Method: MLE Df Model: 41  Date: Fri, 22 Jan 2021 Pseudo R-squ.: inf  Time: 16:45:03 Log-Likelihood: -inf  converged: True LL-Null: 0.0000  Covariance Type: nonrobust LLR p-value: 1.000  =================================================================================================  coef std err z P>|z| [0.025 0.975]  -------------------------------------------------------------------------------------------------  CaseOrder -0.0001 5.61e-05 -1.836 0.066 -0.000 6.95e-06  Zip 2.59e-05 7.5e-06 3.454 0.001 1.12e-05 4.06e-05  Lat -0.0237 0.017 -1.372 0.170 -0.058 0.010  Lng 0.0438 0.014 3.235 0.001 0.017 0.070  Population -7.015e-06 6.61e-06 -1.062 0.288 -2e-05 5.93e-06  Children -0.2865 0.088 -3.242 0.001 -0.460 -0.113  Age 0.0230 0.009 2.467 0.014 0.005 0.041  Income -3.93e-06 2.95e-06 -1.332 0.183 -9.71e-06 1.85e-06  Outage\_sec\_perweek -0.0621 0.029 -2.150 0.032 -0.119 -0.005  Email -0.0372 0.028 -1.341 0.180 -0.092 0.017  Contacts -0.0484 0.090 -0.537 0.591 -0.225 0.128  Yearly\_equip\_failure 0.0586 0.136 0.433 0.665 -0.207 0.324  Tenure -0.7638 0.213 -3.582 0.000 -1.182 -0.346  MonthlyCharge 0.0261 0.008 3.439 0.001 0.011 0.041  Bandwidth\_GB\_Year 0.0080 0.003 3.074 0.002 0.003 0.013  Item1 -0.0295 0.125 -0.235 0.814 -0.275 0.216  Item2 0.1842 0.118 1.565 0.118 -0.047 0.415  Item3 -0.0123 0.109 -0.113 0.910 -0.227 0.202  Item4 -0.0184 0.094 -0.196 0.845 -0.202 0.165  Item5 -0.4313 0.097 -4.434 0.000 -0.622 -0.241  Item6 -0.2401 0.103 -2.323 0.020 -0.443 -0.038  Item7 -0.1691 0.094 -1.801 0.072 -0.353 0.015  Item8 -0.1915 0.093 -2.059 0.039 -0.374 -0.009  PayMethod\_Bank Transfer(automatic) -0.1681 0.267 -0.630 0.528 -0.691 0.354  PayMethod\_Credit Card (automatic) -0.0531 0.259 -0.205 0.838 -0.561 0.455  PayMethod\_Electronic Check 0.2515 0.233 1.080 0.280 -0.205 0.708  Contract\_Month-to-month 3.3333 0.279 11.928 0.000 2.786 3.881  Contract\_One year -0.0914 0.300 -0.305 0.761 -0.679 0.497  Marital\_Divorced -0.3240 0.271 -1.197 0.231 -0.854 0.207  Marital\_Married -0.1505 0.277 -0.543 0.587 -0.694 0.393  Marital\_Never Married -0.0371 0.275 -0.135 0.893 -0.577 0.502  Marital\_Separated 0.2065 0.281 0.735 0.462 -0.344 0.757  Internet\_DSL -1.9964 1.006 -1.984 0.047 -3.969 -0.024  Internet\_Fiber Optic -0.5797 0.353 -1.644 0.100 -1.271 0.111  modem\_Yes 0.1599 0.175 0.915 0.360 -0.183 0.503  phone\_Yes -0.5559 0.280 -1.983 0.047 -1.105 -0.007  security\_Yes -0.8445 0.257 -3.291 0.001 -1.347 -0.342  Backup\_Yes -0.4769 0.234 -2.034 0.042 -0.936 -0.017  Protection\_Yes -0.5082 0.239 -2.126 0.033 -0.977 -0.040  TechSupport\_Yes -0.2825 0.204 -1.386 0.166 -0.682 0.117  StrTV\_Yes -0.2860 0.441 -0.648 0.517 -1.151 0.578  StrMovies\_Yes 0.5468 0.381 1.436 0.151 -0.200 1.293  ================================================================================================= |
| Chart 12 – Statsmodel Logit Summary Data |

Interpretation of this chart focuses on the fourth column of data or the p-value. Any feature with a p-value smaller than 0.05 is a viable candidate for selection as a feature for further investigation.

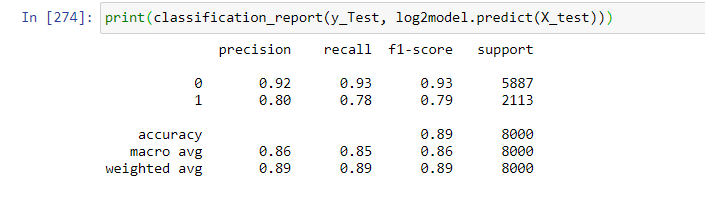
Upon review, Children, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, Item5, Contract\_Month\_to\_Month, Protection\_Yes, Backup\_Yes, Phone\_Yes all fall below the 0.05 threshold.

There are a few extra features notable in this analysis than in the correlation heatmap. However, it does confirm those highlighted by the correlation heatmap.

1. The Reduced Logistic Regression analysis produced improved predictive numbers. While utilizing nine features instead of the full 42 and following the same program as noted in the 'ALL PREDICTORS' model, the scikitLearn LogisticRegression score method posted an 0.88975, a 5.42% improvement. The new confusion matrix for the reduced feature displays the progress with the decreasing numbers of false positives and false negatives.

|  |
| --- |
|  |
| Chart 13 – Confusion Matrix Reduce Predictors |

The new classification report produced these performance numbers for the reduced predictor regression.



Compared to the 'ALL PREDICTOR' model, the prediction value for '1' or customer that 'churn' increased to 79%. That is a 14.5% increase. The recall value, the number of True Positives identified correctly, increased from 65% to 78%, a 20% increase over the 'ALL PREDICTOR' model.

**E. Data Analysis**

**1.** The question we are trying to answer is can we predict churn. Churn is the loss of customers. The idea of churn is that there is something about our service or product that the customer is not delighted in, and they no longer desire to pay for your product. They are probably going somewhere else for these services. So, logically, the parameters relevant to this question would revolve around interactions with the customer. These interactions would be contacts, contracts, Yearly\_equip\_failure, tenure, monthlycharge, Bandwidth\_GB\_Year (how much they use our service), the results from the surveys (items1 through item8), and the types of service they use like Internet service, modem, phone, security, backup, techsupport, and streaming services. Logically, the more company services a customer is using, the more invested they are in the service and the less likely they would change.

To test out that theory, using all the predictors is a general place to start. Evaluating the results and trimming down the features or predictors that do not make an impact. To assess that impact, we utilized the summary section of the statsmodel's Logit function. Each predictor is analyzed in the summary and given a series of statistics. The most important for this evaluation is the p-value noted as P>|z|.

Focus is on the p-values that are 0.05 and below. This value reduces out the insignificant predictors and trims the number from 42 predictors to nine. Using these nine predictors reduces the complexity of the regression and provides better results. All parameters of the regression's predicting capability were improved, including accuracy, precision, and recall.

The model evaluation is based on its predicting capability. These parameters, as noted above, are accuracy, precision, and recall. Accuracy is the correctness of predicting True Negatives and True Positives. Precision is the percentage of True Positive predicted of all Positives predicted both True and False Positives. Recall is the percentage of True Positives predicted of all True Positives.

|  |
| --- |
| ALL PREDICTORS SCORE |
|  |
| Table 1: All Predictors Scoring Results  NOTE: 0 is equal to customers who did not churn, while 1 identifies churned customers. |
|  |
| REDUCED PREDICTORS SCORE |
|  |
| Table 2: Reduced Predictors Scoring Results  NOTE: 0 is equal to customers who did not churn, while 1 identifies churned customers. |

The REDUCED PREDICTORS model scored better on all parameters than the ALL PREDICTORS model. More importantly, the increase in predicting the churn customers was significant at 20% of recall. The importance in that is that in identifying the potential churn customers, The company can take steps that may encourage them to stay as customers.

**2.** Output

**3.** Code See attached file Churn\_Analysis\_Logistic\_Regression.ipynb

**Part V:**

**DATA SUMMARY AND IMPLICATIONS**

**F. Summarize Findings**

**1.Data Analysis**

**Regression Equation:**

y = 1.71(Contract\_Month-to-month + -0.32(Tenure) + -0.24(Children) + -0.34(Protection\_Yes) + -2.43(Phone\_Yes)+ 0.01(MonthlyCharge) + 0.003(Bandwidth\_GB\_Year)

**Interpretation of Coefficients**

The coefficients are a multiplier of the predictor positively or negatively. Regarding the current question, the coefficient will reflect a positive value for a predictor that favors churning, while a negative value is a predictor that favors not churning.

The positive predictors for churn are Contract\_Month-to-month, MonthlyCharge, and Bandwidth\_GB\_Year, while Tenure, Children, Protection\_Yes, Phone\_Yes are opposing forces for churn.

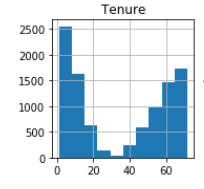
A customer with a Month\_to\_Month contract is the most impactful predictor for churning with a 1.71 multiplier. The Phone\_Yes predictor is the largest predictor of not churning with a -2.43 multiplier. The Protection\_Yes predictor is also a good predictor for not churning with a -0.34 multiplier along with the Tenure predictor with a multiple of -0.32.

**Limitations of Data Analysis**

As with any data analysis, it depends on the dataset is representative of the whole population. The data is clean and free from duplicates and derived values. However, it is just 100,000 records of a larger dataset, reflective of a limited time-frame.

**2. Recommendations**

Churn will inevitably happen. However, reflective in the Tenure predictor distribution, there was some change five years ago that increased the churn rate higher than incoming customers. Even worse, all new customers from 24 to 36 months ago are virtually all gone.



Research into the marketing activities to evaluate the performance and do the opposite.

One expected outcome of the dataset analysis was that the more features a customer would pay for, the less likely they would churn. The data analysis bears that out in a way. It appears that phone, children, and protection services impact customers in staying.

There appear to be no technical glitches that have impacted an area or a product line, and technical support is not an impactful level lacking any significance. There is no connection with technical support calls and churn. The tech support group is doing a good job.

Ideas on how to decrease customer churn? Create incentives for customers to increase the utilization of various services provided. Provide an incentive for customers to sign on for 1-year or 2-year contract terms.

**REFERENCES**

The Art of Statistics How to Learn from Data, David Spiegelhalter

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html>