**D209 Data Mining I**

**Josh Rogers**

**Western Governors University**

**Contents**

**Part I: Research Question 3**

A1. Research question 3

A2. Goals 3

**Part II: Method Justification 3**

B1. Classification Method 3

B2.  K-neighbors Assumption 3

B3.  Packages/Libraries 4

**Part III: Data Preparation 5**

C1.  Data preprocessing goal 5

C2. Initial data set variables 5

C3.  Data Preparation 5

C4.  Cleaned Dataset 7

**Part IV: Analysis 7**

D1. Training and Test Files 7

D2.  Analysis Technique 7

**Part V: Data Summary and Implications 9**

E1.  Accuracy and AUC 9

E2.  Results and Implications 10

E3.  Limitations 10

E4.  Recommendations 10

**Sources 11**

**Part I: Research Question**

**A1. Research Question**

Is it possible to implement the k-nearest neighbors method to identify customers at risk of churning to improve customer retention efforts?

**A2.  Goals**

The goal of this analysis is to build a model utilizing k-nearest neighbors to highlight customers at risk of churning. A feature selection process will also be executed to identify the dependent variables that have the most predictive power. With all this information collected a recommended course of action will be provided to the telecommunication company.

**Part II: Method Justification**

**B1. Classification Method**

The K-Nearest Neighbors classification method is a simple tool that is commonly used due to its ability to be implemented easily. This method, for the purpose of classification, uses the mode of the closest observations for predictive purposes (Hachcham, 2023). A specific k value is selected for the algorithm to identify the number of nearest neighbors to be considered. The algorithm begins by taking a given observation and calculating all distances from all other points. Then, it utilizes the smallest distances of k number of datapoints and calculates the mode.

While this methodology has several advantages it does have its flaws. For example, K-Nearest Neighbors does not scale well due to its large computing and storage requirements. Additionally, the curse of dimensionality is a potential pitfall programmers may experience with this classification method. That is to say, the inclusion of too many features that results in crossing the threshold of a highly predictive model to one that is not as effective (K-nearest Neighbor(KNN) algorithm, 2023). Due to the issue associated with continuous feature additions this method does also tend to overfit. However, in this model, LASSO feature selection is going to be implemented to address this issue.

An expected outcome of this analysis is that the features will be reduced significantly as a result of the LASSO feature selection. Second, a model with an accuracy of over 80% is expected to be developed by using the KNN classification method.

**B2.  K-Nearest Neighbor Assumption**

The K-Nearest Neighbors algorithm assumes that there is similarity of things that exist in close proximity to each other.

**B3. Packages/Libraries**

A screenshot of a computer

Description automatically generated

The **pandas** library was utilized to call in the CSV file and create a pandas data frame which was cleaned and prepared for analysis. This includes the implementation of the getdummies() function to convert the categorical variables to numerical values. This library was also called to convert the split results to data frames for conversion to CSV files.

**Seaborn** and **matplotlib** were imported to assist in displaying the LASSO coefficients to determine which features to include in the algorithm. Additionally, these libraries were called in conjunction with the **sklearn.metrics** library and importing **RocCurveDisplay** to visualize the ROC curve and the AUC calculation.

Feature reduction in this model was performed by importing the **Lasso** function from **sklearn.linear\_model** library. The **StandardScaler** function was imported from **sklearn.preprocessing** to standardize the numeric dependent variables.

The model was constructed via the **KNeighborsClassifier** housed in the **sklearn.neighbors library**. **KNeighborsClassifier** was also referenced to optimize the k value for the final model by printing the accuracy scores for k values 1-15. To train the model with the selected features from the **Lasso** function, the **train\_test\_split** function was imported from the **sklearn.model\_selection** library.

**Part III: Data Preparation**

**C1.  Describe one data preprocessing goal.**

To create a more accurate model the Lasso feature selection method was implemented. This regularization method was used to highlight features that were less important and thus eliminated from the final model. This step is an important part of the process in developing a K-Nearest Neighbors algorithm especially due to the potential inclusion of too many variables which can reduce model efficiency.

**C2. Identify the initial data set variables that you will use to perform the analysis for the classification question from part A1 and classify each variable as continuous or categorical.**

Below is a list of the initial variables being considered to answer the research question along with their associated variable type prior to feature selection:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** |  |
| Churn | Categorical |  |
| MonthlyCharge | Numeric |  |
| Tenure | Numeric |  |
| Children | Numeric |  |
| Age | Numeric |  |
| Income | Numeric |  |
| Gender\_Male | Categorical |  |
| Gender\_Female | Categorical |  |
| Gender\_Nonbinary | Categorical |  |
| Contract\_Month-to-month | Categorical |  |
| Contract\_One year | Categorical |  |
| Contract\_Two Year | Categorical |  |
| PaymentMethod\_Bank Transfer(automatic) | Categorical |  |
| PaymentMethod\_Credit card (automatic) | Categorical |  |
| PaymentMethod\_electronic Check | Categorical |  |
| PaymentMethod\_Mailed Check | Categorical |  |

**C3. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.**

Multiple actions were taken in wrangling the data to build the required models. First, when the CSV file was read in via Pandas, only the columns housing the selected explanatory variables along with the dependent variable (Churn) were pulled into the analysis. After this the data was filtered to remove null values by utilizing the dropna() function. (Code segment below)

A screenshot of a computer

Description automatically generated

After these initial steps it was necessary to adjust the categorical variables to numerical values. First these columns were identified and stored in a pandas data frame called ‘categorical\_cols’. Then, by implementing the getdummies() function, the ‘categorical\_cols’ were passed through. Removal of the first output was set to ‘False’ and the datatype was set to output an integer for ease of calculations moving forward. (Code segment below)

A screenshot of a computer

Description automatically generated

It was necessary to standardize the numerical feature variables in preparation for the model construction. To do so, the numeric columns were identified and the StandardScaler function was utilized to fit and transform this data. (Code segment below)

A screenshot of a computer

Description automatically generated

The categorical values were then concatenated with the original filtered data and the original categorical column names were dropped to avoid confusion with those that were generated by the getdummies() function. The ‘Churn\_Yes’ column was renamed to ‘Churn’ due to programmer preference. (Code segment below)

A screenshot of a computer screen

Description automatically generated

**C4. Provide a copy of the cleaned data set.**

Attached**: D209\_T1\_Churn\_Dataset.csv**

**Part IV: Analysis**

**D1.  Split the data into training and test data sets and provide the file(s)**

The following files have been submitted separate from this Word document:

X\_test.csv

X\_train.csv

y\_test.csv

y\_train.csv

**D2.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.**

Below is the code segment and a screenshot of the visualization of the results from the Lasso feature elimination process. It was determined that the elimination of all variables except Tenure and MonthlyCharge would optimize the KNN algorithm. Once the features were identified the x and y values were defined with only those that were to be used moving forward. Below is the code segment for these operations as well as the visualization of the Lasso coefficients which was used to constrict the model features.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

To further optimize the KNN algorithm an analysis was performed to determine the best k-value accuracy to be included. The code segment and output below show that the accuracy score for k-values 1-15 were considered and ultimately the k-value = 13 was selected for inclusion in the algorithm.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Part V: Data Summary and Implications**

**E1.  Explain the accuracy and the area under the curve (AUC) of your classification model.**

To analyze the performance of the algorithm a visualization was created to display the ROC-AUC curve and calculated AUC value. Below it can be seen that the AUC = .89. As a value closer to 1 is associated with better performance the resulting AUC value suggests the algorithm performs quite well. This means that the model has an 89% chance to predict customer churn based on the feature variables of MonthlyCharge and Tenure.

A screenshot of a computer

Description automatically generated

**E2.  Discuss the results and implications of your classification analysis.**

Upon review it has been determined that the constructed algorithm performed well with an AUC value of 0.89. However, given that there were only two features selected there is a possibility that additional features, not included for consideration in the onset of this analysis, could have been important in bolstering the algorithm’s effectiveness.

Additionally, while the Lasso feature selection methodology was utilized on the specifically selected features, the model may have had better performance utilizing the SelectKBest function. By using this function, the algorithm could have included more features from all possible options in the dataset.

**E3. Discuss one limitation of your data analysis.**

The dataset being utilized for this model includes 10,000 customer accounts with 2,650 customers with a churn value of “Yes” and 7,350 of a value of “No”. This imbalance could result in the training dataset to focus too much on those that did not churn due to its higher likelihood of accurately predicting this class (Introduction to balanced and Imbalanced datasets in machine learning, n.d.). In the business scenario we are considering the focus is intended to be on how to better retain those that are more likely to churn thus the imbalance of the dataset may hamper real world application of this model.

**E4. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.**

While the reduced algorithm yielded positive results in its predictive power the business applications must be used with caution. Due to the imbalanced nature of the dataset this model should not be leaned on as the only analytical tool for decision making. However, business decisions can cautiously be made regarding the two feature variables that were selected by the reduction method for this model, MonthlyCharge and Tenure. For example, marketing efforts focused on those with lower tenure and high monthly charges can be developed to promote greater customer longevity and thus reducing the churn percentage.

**Sources**

**H. Code Sources**

*Python KNN: Mastering K nearest neighbor regression with sklearn*. (n.d.). Welcome to Kanaries Docs – Kanaries. https://docs.kanaries.net/topics/Python/python-knn#:~:text=To%20use%20the%20KNeighborsRegressor%2C%20we%20first%20import%20it%3A,our%20model%20to%20the%20data%20and%20make%20predictions%3A

**I. Content Sources**

Filho, M. (2023) *How to solve logistic regression not converging in Scikit-learn*, *Forecastegy (Alt + H)*. Available at: https://forecastegy.com/posts/how-to-solve-logistic-regression-not-converging-in-scikit-learn/ (Accessed: 13 March 2024).

Hachcham, A. (2023, August 11). The KNN algorithm - Explanation, opportunities, limitations. neptune.ai. https://neptune.ai/blog/knn-algorithm-explanation-opportunities-limitations

*Introduction to balanced and Imbalanced datasets in machine learning*. (n.d.). Unlock AI From Your Data | Encord. https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/

*K-nearest Neighbor(KNN) algorithm*. (2023, November 9). GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/

*Logistic regression calculator*. (n.d.). Online Statistics Suite. https://stats.blue/Stats\_Suite/logistic\_regression\_calculator.html

*Logistic Regression in Machine Learning*. (2017, May 9). GeeksforGeeks. https://www.geeksforgeeks.org/understanding-logistic-regression/#assumptions-of-logistic-regression

R or Python. (2023, July 7). Retrieved from https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

Tewari, U. (2021, November 9). Regularization — Understanding L1 and L2 regularization for Deep Learning. Medium.com. Retrieved March 14, 2024, from https://medium.com/analytics-vidhya/regularization-understanding-l1-and-l2-regularization-for-deep-learning-a7b9e4a409bf

Tim (2020) *Regularization: Simple definition, L1 & L2 penalties*, *Statistics How To*. Available at: https://www.statisticshow