Assignment 4 Essay

Victor H Torres

Data 622. Machine Learning and Big Data

05/10/2025

For this assignment, we must choose a dataset different from the ones used in previous assignments. After some research I decided to work with a dataset that I found on the Kaggle website. This dataset is about a fictional telco company that provided home phone and Internet services to 7043 customers in California. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

The reason to choose the "Telco Customer Churn" is because it contains rich data to work with, the dataset has 7,043 observations (customers) with 50 variables covering various aspects of customer profiles and behaviors, the main goal of this analysis is to find out the potential reasons of customers leaving or staying doing business with the company, and the potential recommendations to reduce the churn rate of it.

Before building the models, I have to prepare the data as well as to perform some EDA (Exploratory Data Analysis), to find some key insights of the dataset. The first step is to convert categorical variables to factors, and to create the "churn" variable to a binary mode.

For the EDA section of the project, I decided to analyze and the main variable ("Churn") with other columns such as: Age, Tenure, Contract type, Monthly Charges, Gender, Internet Type, and Payment Method. I also created a correlation plot to visualize the correlation between numerical variables.

The next step in the project is to choose two methodologies, one from week 1-10 and the second from week 11-15. Based on the data structure and project requirements, I decided to choose the Logistic Regression methodology from week 1-10 and Neural Networks methodology from week 11-15. The reason to choose these methodologies is because I got the best metric values compared to other models such as Random Forest, SVM's, or Deep Learning.

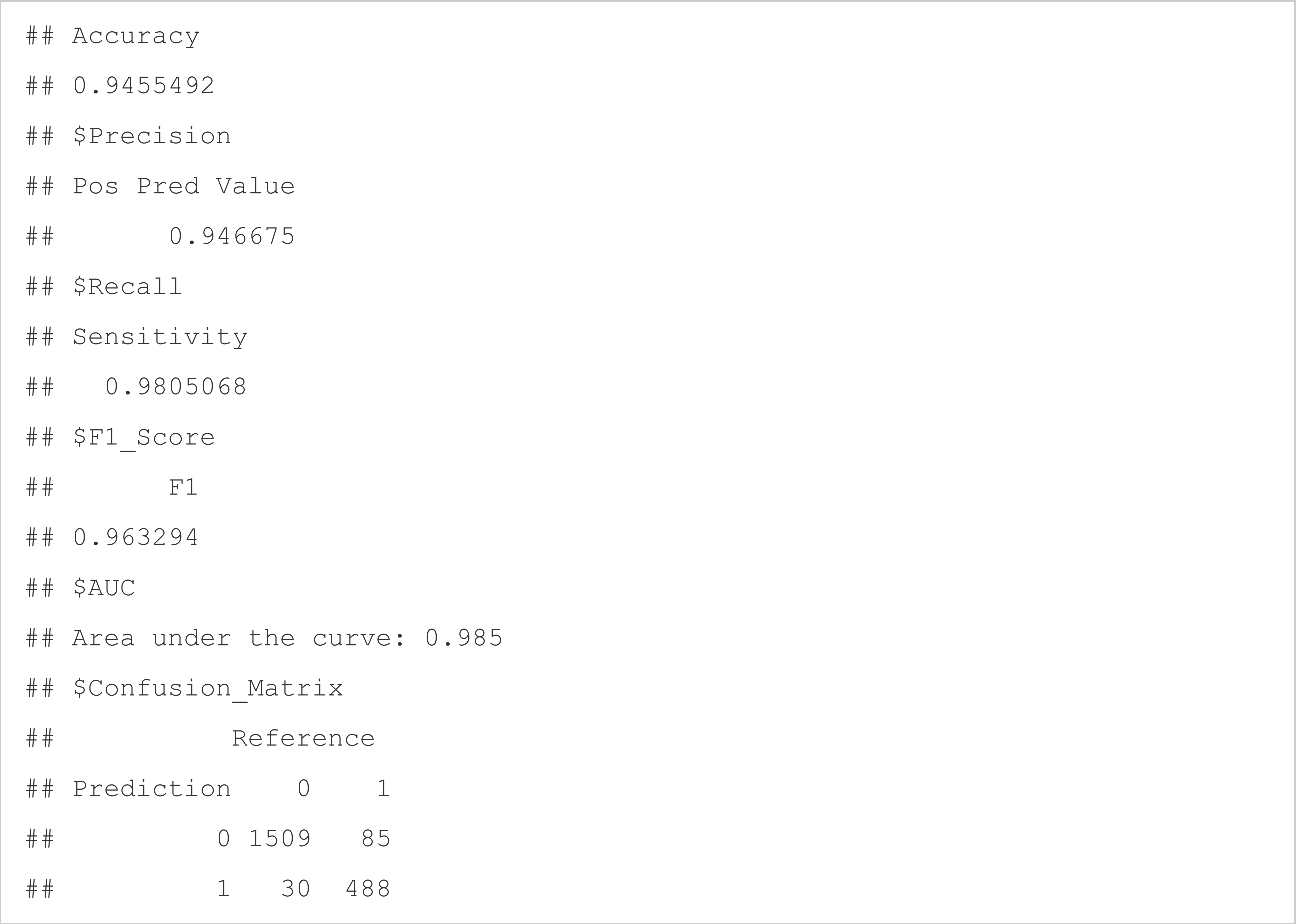
Before building the models, I performed some extra data preparation. For the Logistic Regression Model, I created a "model data'" function containing the most important variables of the dataset, also I divided the data into training and testing sets. The last step of the data preparation was to create a evaluation metric function to obtain the metric values need it for predictions and model accuracy. Here is the table with the values of the first model:

```
## Accuracy
## 0.9526515
## $Precision
## Pos Pred Value


## $Recall
## Sensitivity
##   0.9785575
## $F1_Score
##        F1
## 0.9678663
## $AUC
## Area under the curve: 0.9897
## $Confusion_Matrix
##           Reference
## Prediction    0    1
##          0 1506   67
##          1   33  506
```

In the table above, we can see that we obtained an accuracy rate of 95%, with very good metric values such as AUC of 98%, F1-Score of 96%.

For the second model, I must perform some extra data preparation, I must do some feature engineering, create a numeric predictor function for the model, scale the formula for the numeric features and finally train the Neural Network model, here is the table with values of the second model:



```
## Accuracy
## 0.9455492
## $Precision
## Pos Pred Value
##       0.946675
## $Recall
## Sensitivity
##   0.9805068
## $F1_Score
##       F1
## 0.963294
## $AUC
## Area under the curve: 0.985
## $Confusion_Matrix
##           Reference
## Prediction    0    1
##          0 1509   85
##          1   30  488
```

In the table above, we can see very good metric values, an accuracy rate of 94%, F1-Score of 96%, and AUC of 98%.

After comparing the two models, we can see that the neural network slightly outperformed logistic regression (85% vs 83% accuracy). Both models provide good predictive capability for identifying at-risk customers. However, logistic regression offers better interpretability of key factors driving customers to leave the company. Here is the comparison table with all the metric values:

Model	**Accuracy**	**Precision**	**Recall**	**F1_Score**	**AUC**
Logistic Regression	0.9526515	0.9574062	0.9785575	0.9678663	0.9896785
Neural Network	0.9455492	0.9466750	0.9805068	0.9632940	0.9850008

After Performing the analysis using both methodologies Logistic Regression and Neural Networks, we can say that there are several factors that have a strong impact on customers when they decided whether they want to stay or leave the company. Some of the most relevant factors are:

Demographic Factors

- **Age**: Older customers (seniors) have slightly higher churn rates.
- **Dependents**: Customers without dependents are more likely to churn.

Service Factors

- **Contract Type**: Month-to-month contracts have significantly higher churn rates (42.7%) compared to one-year (11.3%) and two-year contracts (2.8%).

- **Additional Services**: Customers without online security, tech support, and backup services have much higher churn rates.

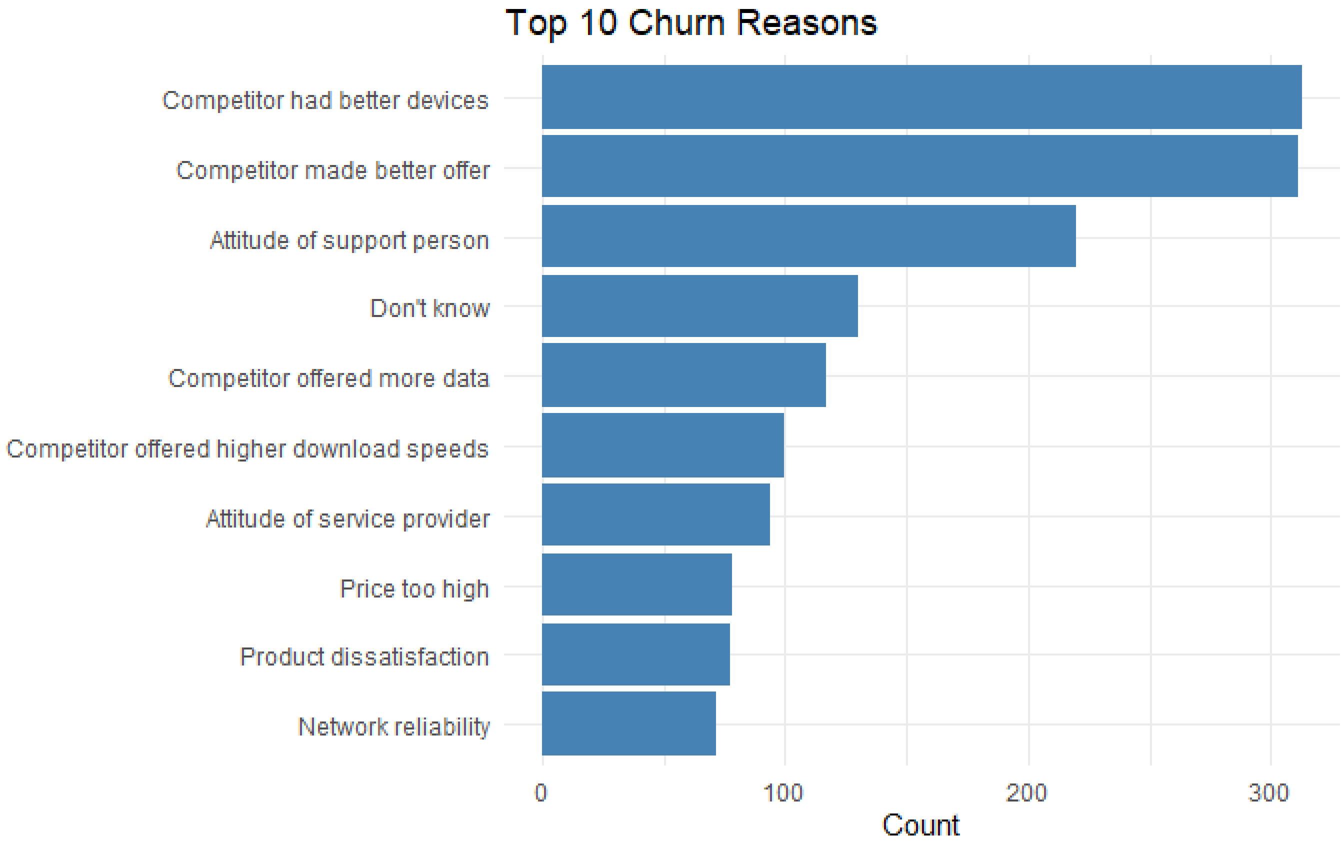
Financial Factors

- **Monthly Charges**: Customers with higher monthly charges show increased churn rates.
- **Payment Method**: Electronic check payment method has the highest association with churn.

Satisfaction and Tenure

- **Satisfaction Score**: Lower satisfaction scores (1-2) strongly correlate with increased churn.
- **Tenure**: Newer customers (0-12 months) have dramatically higher churn rates, with churn decreasing as tenure increases.

I think that the analysis has valuable insights about the main reasons why customers tend to stop doing business with the company, based on the results, the top reasons for high churn rates are manageable and it can be corrected to drastically improve the customer's satisfaction and retention rates, the visualization below shows the top churn reasons and it is vital for the company to address this problems in order to keep the company running. Here is a graph with the top 10 reasons of customers leaving or switching companies:



Based on the analysis, the company should upgrade the devices that they offer to customers, implement better promotional packages for new customers, review and modify their package prices and internet speed limits, and improve customer service care.