**Srinivas Bhootam**

**DeVos Graduate School, Northwood University**

**115537-MGT-665-NW Solv Probs W/ Machine Learning**

**Customer Churn Prediction: Logistic Regression, k-NN, and Decision Tree**

**Dr. Itauma Itauma**

**June 15th, 2025.**

**Abstract:**  
 In this research supervised machine learning algorithms have been utilized for predicting the Customer Churn for the Telco Customer Churn dataset. Churn, when customers desert one service, is a key problem for telecom companies. I will Construct and compare performance metrics of following 3 classification models they are Logistic Regression, k-NN and Decision Tree. I will preprocess, visualize, and analyze the dataset before training and evaluating models. In the three models, the Logistic Regression is the model with the most balanced accuracy across the measures. The paper provides a discussion on the steps of model development, insights into evaluation, and the outlook for which improvements could be introduced. Such in-sights can be used by telecommunication companies to report and forecast churn to minimize its impact on profitability.

**Introduction:**

Customer churn prediction is particularly useful for companies that provide recurring services, such as telecom companies, since retaining customers is often less expensive than acquiring new ones. Once a company knows which of its customers are most likely to leave, it can take steps to keep those customers by offering them better deals or posture with better service. In this paper I will develop and compare three machine learning models – Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree, using actual customer data set from a telecom company. These models allow us to discern patterns in the data that indicate what customers are most likely to churn. Also compare how well each model does using basic metrics like accuracy, precision, recall, and F1-score and then decide which model is ideal for predicting churn and assisting companies to minimize customer attrition.

**Literature review:**

Customer churn has been well researched for years; it’s a significant problem for firms who depend on repeat business. Indeed, some researchers such as Hadden et al. (2007) demonstrated that computerized decision making can improve the ability of companies to learn about why customers take their business elsewhere and respond appropriately. Logistic Regression is a common model used for this purpose since it is easy to use and reveals what increases the odds of a customer leaving (Han, Kamber, & Pei, 2011).

Other models such as k-Nearest Neighbors (k-NN) and Decision Trees are also utilized. k-NN is easy to use but does not perform well if the data is unbalanced or not properly scaled (Ahmed et al., 2016). Decision Trees can manage a wide range of different types of data, providing simple and easy-to-understand decision rules, while at the same time they can overfit and not generalize properly for new data (Pedregosa et al., 2011). Though more recent approaches such as deep learning make little sense economically, these simpler models remain valuable because they are fast, easy to understand, and effective in numerous commercial applications.

**Methodology:**

The Telco Customer Churn dataset consists of the following 7,043 rows and 21 columns, which contain information about multiple aspects of the customer such as their demographic information, the services that they have subscribed to, contract type, and monthly charges: The response variable, Churn, is a customer becoming not being a customer.

Pre-analysis Data processing: CustomerID column was removed as it doesn’t contribute towards prediction. The TotalCharges column had empty strings or strings entries, which were changed to numerics and any NA’s deleted. Categorical values were converted to numerical values using labels encoding them to make them applicable to machine learning.

Once the data was cleaned and encoded, it was split into input features (X) and the target variable (y). 80% and 20% division ratio was used for splitting the data into training and testing sets. As certain algorithms (e.g. Logistic Regression, k-Nearest Neighbors) are sensitive to feature scaling, we scaled the numeric columns with StandardScaler. This step was taken to make sure that variables such as tenure and MonthlyCharges contributed equally in the learning.

**Model Description:**

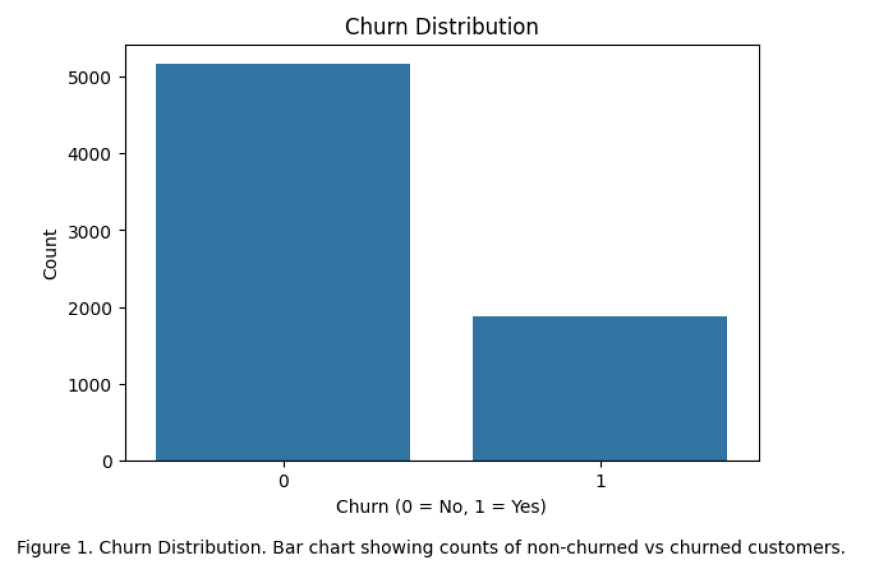
In this paper, three supervised machines learning algorithms such as Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree Classifier were used to predict customer churn. We chose these two models, because they are universally used for classification tasks, simple to implement, and interpretable.

Logistic Regression (LR). Also known as Logit Regression, LR is a statistical method for analyzing a data sets in which there are one or more independent variables that determine an outcome: in this example if a customer will or will not churn. It models the relationship between the dependent and one or more independent variable using a logistic (sigmoid) function. This model is especially convenient to interpret when it comes to the importance and direction of impact of each characteristic on the probability of the churn.

k-Nearest Neighbors (k-NN) is a simple non-parametric, instance-based learning algorithm that labels a new observation according to the majority class of its nearest neighbors. In this study, the k value was 5, so the prediction was calculated using the five most similar data in the training dataset. Simple though it is, the model performs subject to feature rescaling and data imbalance.

Decision Tree Classifier is another model based on decision trees that cuts recursively a set of data in subsets with a decision rule taking each subset. Every internal node is a decision on a feature, while each leaf node is a class label prediction. Even though the decision trees are highly interpretable, but with low level models they have the same issue of overfitting on the (small) dataset, unless you have set high value for appropriate parameters.

**Results**

****

**A graph of a number of numbers

AI-generated content may be incorrect.**

**A diagram of a chart

AI-generated content may be incorrect.**

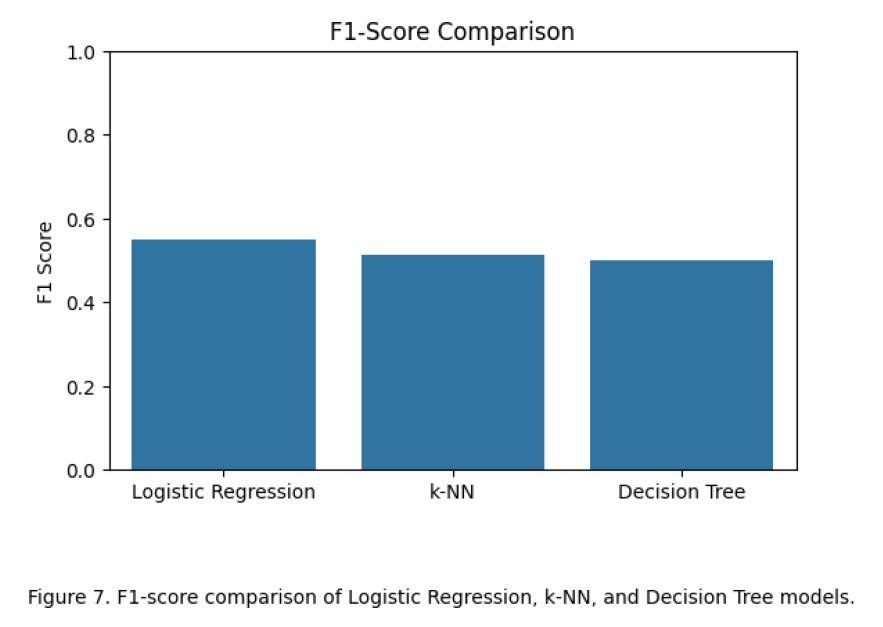
**A screenshot of a computer screen

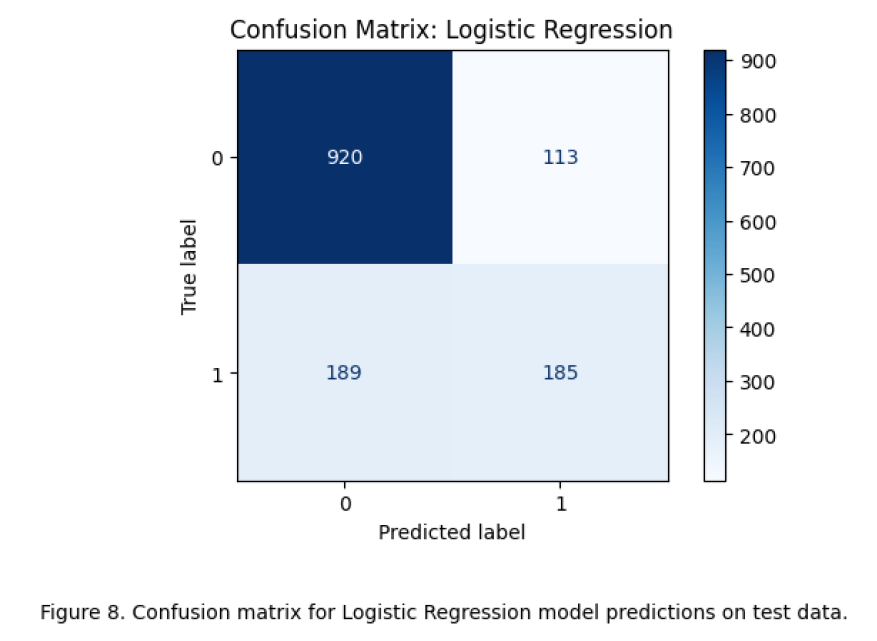
AI-generated content may be incorrect.A graph of a diagram

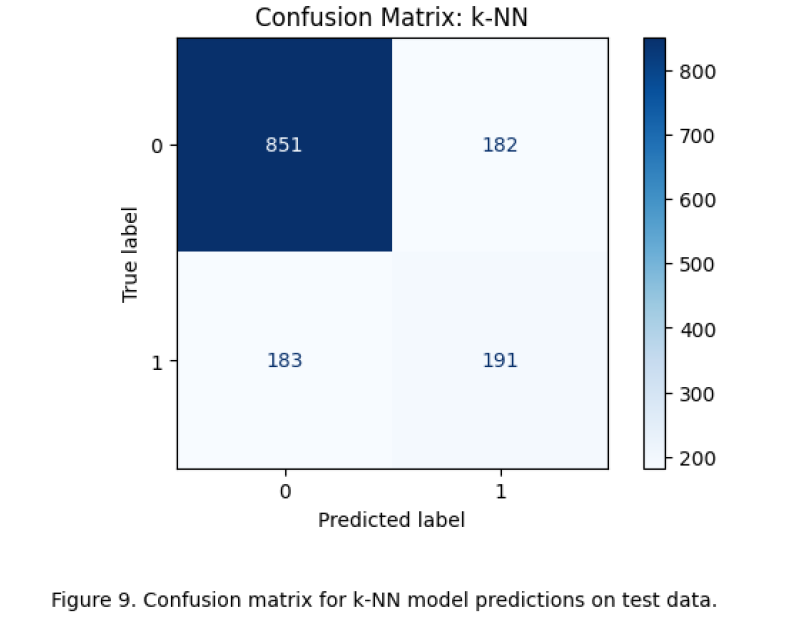
AI-generated content may be incorrect.**

**A screen shot of a graph

AI-generated content may be incorrect.**

****

****

****

**A blue and white graph

AI-generated content may be incorrect.**

**Model Development and Evaluation Metrics:**

I trained and tested three models: Logistic Regression, k-Nearest Neighbors (k=5), and Decision Tree Classifier. In all cases, models were trained on the training data set and applied to the same test data set for concordance.

* The models were tested in terms of:
* Precision is the fraction of true hits among all predicted positive instances.
* Precision: This measure tells us what proportion of customers that we targeted to churn cases did churn.
* Recall is checking how well the model identifies actual churn customers.
* F1-score integrates both precision and recall into a single metric, which is critical when the number of negative examples is largely unbalanced to the positive ones.

To gain deeper insight about the correct vs. incorrect prediction distribution, confusion matrices were consulted as well (Figures 8–10). A correlation heatmap (Fig. 6) and class distribution chart (Fig. 5) were useful to interpret the feature-target relationships. Furthermore, Figure 7 visualizes the spread of F1-scores for all implemented models.

**Discussion:**

In general, the best performance was given by the Logistic Regression model. It had a 79 percent accuracy and an F1-score of 0.55, so it did a good job in correctly identifying both who stayed and who left. Its precision and recall were close to each other, indicating that it was not biased toward either class. It also made fewer errors compared to the other models as evidenced in confusion matrix (Figure 8). Due to its stable and reliable performance, Logistic Regression was the most robust model in the present research.

The k-Nearest Neighbors (k-NN) model resulted in 74% accuracy and an F1-score of 0.51. It was significantly more sensitive to problems such as not carrying the same number of samples in each class and the scale of the data. This had the effect of making it more difficult for the model to distinguish customers who would churn. The Decision Tree model had the highest recall (0.52), which implies this model was the most capable of finding churned customers. But it was less precise (0.48), so it also incorrectly guessed that that many loyal customers would leave. This could result in ill-advised responses from the business. The confusion matrices in Figures 9 and 10, as well as class imbalance in Fig. 5 show that it is imperative to consider other metrics than accuracy when evaluating models.

**Conclusion:**

In this paper, the performance of three machine learning methods-Logistic Regression, k-NN, and Decision Tree were compared in terms of customer churn prediction based on the Telco dataset. All models were constructed with Python’s scikit-learn library and conforming the pre-processing and evaluation procedures.

Logistic Regression is the best performing method out of three, with it a good balance of precision and recall as well as computationally fast. The Decision Tree model proved to be promising with great recall, which will be ideal for businesses where missed churners are considered more deleterious than false positives. k-NN, despite being the simplest of architecture, had limitations given the features of the present dataset.

For future development, we suggest exploring ensemble models such as RandomForest or XGBoost and class imbalance corrective methods such as SMOTE to mitigate the effects of the negative class imbalance. The robustness of the model may be increased by adding the time-series features or behavioral data. By optimizing these models, telecom operators can proactively predict and retain customers before they are churning and save a significant amount on churn-related losses.

**References**

Ahmed, S., Ameen, A., & Rizwan, M. (2016). Predicting Customer Churn in Telecom Industry Using Multilayer Perceptron Neural Networks: A Case of Ethiopia. International Journal of Computer Applications, 975, 8887.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902–2917.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.