# Assessment 1.2: Feature Selection for Customer Churn Prediction: Exploring Genetic Algorithms and Crossover Strategies

**Sachin, ID: 23235298** [*1]

Word Count: 2175

## Abstract

In the dynamic telecommunications industry, the ability to predict and manage customer churn is paramount for sustaining competitiveness and fostering loyalty. This research delves into churn prediction, specifically within telecommunications, leveraging a fusion of Genetic Algorithms (GAs) and Logistic Regression on historical data. With a focus on feature selection, the study investigates three distinct crossover methodologies—Single-Point, Two-Point, and Uniform—within the GA framework. By optimizing feature selection through these crossovers, the approach enhances the accuracy of churn prediction models. The insights derived empower telecom companies to proactively identify and retain at-risk customers, ultimately enhancing customer satisfaction and long-term retention strategies. [1].

## 1. Introduction

In today's fiercely competitive telecommunications industry, customer churn—where customers switch to competitors—presents a significant challenge. This paper focuses on predicting churn, a crucial aspect of retaining customers and sustaining business growth. By leveraging advanced techniques like Genetic Algorithms and Logistic Regression, our aim is to identify customers who are at risk of churning by analyzing their past behavior and interactions. Our approach incorporates three distinct crossover strategies—Single-Point, Two-Point, and Uniform—to optimize the selection of features used for churn prediction. Through enhancing the accuracy and effectiveness of our predictive model, we provide telecom companies with valuable insights and tools to proactively address customer attrition. Ultimately, this research contributes to fulfilling a pressing need within the industry by empowering telecom companies with actionable strategies to retain their customer base and thrive in the competitive market landscape (Babatunde et al., 2014).

This report is organized as follows: Section 2 introduces the problem domain and relevant literature, setting the context for the research. Section 3 narrows down to the specific problem instance targeted, detailing the dataset and its relevance. In Section 4, candidate optimization methods and chromosome design are presented as solutions. Methodology for experimentation is outlined in Section 5, with results showcased in Section 6, followed by a discussion in Section 7. Finally, Section 8 contains closing thoughts and future research recommendations, wrapping up the study.

## 2. Introduction to the Problem Domain

Churn in the business domain occurs when a customer leaves a service and switches to a competitive provider. Nowadays, customers have many options to choose the best services from, leading to a competitive market such as the telecom industry where loyalty becomes precious.

Attracting new customers in such markets is eminently challenging, and costs five to six times more than preventing existing customers from churning (Stripling et al., 2018). What if businesses can predict the churners? But detecting the churners out of millions of customers isn't an easy task.

Telecommunications companies are experiencing increasing pressure in customer support as they offer bundled Audio, Video, and Internet access services known as the triple play package. The challenge lies in ensuring a high-quality experience for consumers during service usage and when seeking assistance from their providers. With the telecom industry witnessing significant growth in recent years and nearly universal adoption of telecom packages, this paper primarily focuses on this sector. Churn prediction models analyze historical business data to pinpoint clients

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: Sachin <sachin.-3@mail.bcu.ac.uk>.

*Modern Optimisation Report*, CMP 7213 - Modern Optimisation.

[1]https://www.bcu.ac.uk/about-us/education-development-service/centre-for-academic-success-student/resources

at high risk of leaving, enabling telecom companies to focus on specific customer segments rather than individuals. Personalized retention efforts are challenging due to the large customer base and limited time and resources available for investment. (Fujo et al., 2022).

## 3. Problem Instance

The dataset utilized for this research is the IBM Telco dataset, which encompasses diverse attributes pertinent to customers within the telecom industry. This dataset facilitates a deeper comprehension of the correlation between customer actions and churn rates. It comprises 21 attributes and 7043 rows and is readily accessible on Kaggle. This dataset is used by many researchers (Momin et al., 2020),(Amin et al., 2019), (Agrawal et al., 2018),(Mohammad et al., 2019), (Pamina et al., 2019).

The IBM Telco dataset features details about the services each customer has subscribed to, including phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies. Moreover, it comprises details about customer account specifics like their duration as a customer, type of contract, method of payment, preference for paperless billing, monthly charges, and total charges. Demographic information about customers, including gender, age range, and whether they have partners and dependents is also included. Finally, the dataset features a dependent attribute denoted as "Churn," indicating whether a customer has churned or not. Table 1 delineates the dataset's characteristics.

| Characteristics | IBM Telco Dataset |
|---|---|
| Total No. of features | 21 |
| Total No. of customers | 7043 |
| Missing value | Yes |
| Churn | 26.5% |
| Not churn | 73.5% |
| Data distribution | Imbalanced |
| Categorical features | 17 |
| Numerical features | 4 |
| Dependent feature | 1 |
| Independent features | 20 |

*Table 1.* Characteristics of dataset

Many issues in the IBM Telco dataset need attention, such as missing values, non-numeric features, etc. Additionally, the issue of an imbalanced dataset is present. Therefore, it is essential to pre-process the data before implementing a learning model.

The paper (Liu et al., 2007) discusses the significance of imbalanced datasets across various areas of data mining and outlines the assessment metrics and existing techniques for tackling the imbalance issue. Among these methods is the Synthetic Minority Oversampling Technique (SMOTE), which is a form of oversampling designed to deal with this problem. In this research, the SMOTE method has been used to handle imbalanced datasets. A pre-processed dataset helps to achieve better results and attain the objective function, which is maximizing accuracy.

***Objective Function=Maximize (Prediction Accuracy)***

## 4. Candidate Optimization Methods

This section delves into various optimization techniques, notably, Genetic Algorithm (GA) and three types of crossovers—Single-point, Two-point, and Uniform—paired with Logistic Regression for Customer Churn Prediction. Evaluating adaptability and performance is vital in this complex process. The research aims to bridge this gap by comparing the strengths and limitations of these optimization methods, shedding light on their effectiveness in addressing the challenges of churn prediction within the telecommunications industry.

### 4.1. Genetic Algorithm

The Genetic Algorithm (GA) is a metaheuristic optimization method inspired by natural selection principles, particularly the concept of "survival of the fittest." It operates through an iterative process, starting with an initial set of solutions known as the population. Each solution is defined by a set of properties, akin to genes. Following this initialization phase, the optimization objective, often referred to as the fitness function, is evaluated across the population. The fittest solutions are then selected through a process called selection, and their properties are combined to generate a new population through crossover. To prevent the algorithm from converging towards local minima, random changes in properties, known as mutation, are introduced between generations. This process of selection, crossover, and mutation continues iteratively until either computational limits are reached or the optimization objectives defined by the fitness function are achieved (Fridrich, 2017).

### 4.2. Chromosome Design

The chromosome serves a crucial role in feature selection for predicting customer churn using logistic regression. It acts as a binary representation where '1' denotes the inclusion and '0' signifies exclusion of specific features. This tailored combination of features, optimized within the chromosome, is instrumental in effectively identifying churn patterns. The precision in feature selection highlighted by this chromosome underscores its significance in enhancing the predictive power of the logistic regression

model, thereby aiding telecom companies in proactively addressing customer attrition.

**Chromosome = 01011011001111111010110010**

### 4.3. Single-Point Crossover

Single-point crossover in genetic algorithms selects a random point on parent chromosomes, then swaps segments to produce offspring, promoting diversity by combining parental traits. Its effectiveness depends on problem complexity and genetic representation, vital for crafting efficient algorithms tailored to specific applications. This method facilitates exploration of solution space but may struggle with complex problems or require adjustments to genetic representation to optimize performance, illustrating the importance of careful consideration in algorithm design (Abualigah & Dulaimi, 2021).


*Figure 1.* Genetic Exchange: One-Point Crossover

### 4.4. Two-Point Crossover

Two-point crossover in genetic algorithms selects two random points in parent chromosomes, swapping genes between these points to create offspring. It facilitates exploration of solution space by combining genetic information from parents to generate potentially improved offspring. This process is pivotal in problems where gene order is significant, like sequence optimization or scheduling. Through iteratively replacing less fit individuals with offspring, genetic algorithms converge towards optimal solutions by leveraging the recombination of genetic material across generations.(Xue et al., 2021).


*Figure 2.* Genetic Exchange: Two-Point Crossover

### 4.5. Uniform Crossover

Uniform crossover, a genetic algorithm operator, assigns equal probability for each gene in offspring to inherit from either parent, fostering diversity by extensively mixing genetic material. Compared to single-point crossover, it explores a broader solution space. Its efficacy hinges on problem traits and parameter fine-tuning for optimal results. This method offers a versatile approach to solution exploration but requires careful consideration of problem characteristics and parameter settings for effectiveness (Abualigah & Dulaimi, 2021).


*Figure 3.* Genetic Exchange: Uniform Crossover

### 4.6. Logistic Regression

Logistic regression stands out in machine learning for its transparency compared to complex ”black box” models. It can be binary, multinomial, or ordinal, but our focus is on binary logistic regression. Operating with real-valued inputs, it predicts whether they belong to class 0. If the prediction surpasses 0.5, it’s categorized as class 0 (non-churners); otherwise, it’s class 1 (churners).

Logistic regression computes the log odds of churn probability, always constrained within the range of 0 to 1, offering a clear and interpretable framework for understanding and predicting binary outcomes like customer churn (Jain et al., 2020).

## 5. Experimental Setup

This study examines telecom customer churn prediction using logistic regression, leveraging Kaggle-sourced data with varied attributes. Feature selection employs Genetic Algorithm techniques including Single-Point, Two-Point, and Uniform Crossovers. Logistic regression models are implemented in R, with evaluation metrics used to assess predictive performance. The study’s core objective is to achieve accurate churn prediction, crucial for telecom companies in retaining customers and reducing attrition rates. Through this analysis, the research aims to offer valuable insights into the effectiveness of logistic regression models and the impact of different feature selection strategies on predictive accuracy in the telecommunications industry.

### 5.1. Data Description

The IBM Telco dataset sourced from Kaggle, comprising 21 attributes and 7043 rows, facilitates insights into telecom customer churn. Preprocessing involved cleaning the data, converting data types, Synthetic Minority Oversampling Technique to handle imbalanced dataset and employing one-hot encoding for categorical columns. These steps ensure data readiness for analysis and model training.

By preparing the dataset in this manner, researchers can effectively explore patterns and factors influencing customer churn, leading to informed decision-making in the telecommunications industry.

### 5.2. GA Parameters

Decisions regarding population size, crossover techniques, and mutation strategies in genetic algorithms were guided by thorough investigation into optimal parameter selections. Restricting the number of generations to 20 was necessitated by the significant computational resources required for this problem. These choices reflect a balance between achieving effective optimization within computational constraints. Through this approach, the study aims to maximize the efficiency of genetic algorithms in addressing the complexities of telecom customer churn prediction while managing computational costs (Zaharie, 2009).
Table 2 provides a concise overview of parameters crucial for the genetic algorithm, encompassing generation count, population size, crossover and mutation probabilities, data type, crossover techniques, and dataset utilized in the analysis.

| Parameter | Value |
|---|---|
| maxGenerations | 20 |
| popSize | 50 |
| pcrossover | 0.8 |
| pmutation | 0.1 |
| type | binary |
| crossover | Single-Point, Two-Point, Uniform |
| data | IBM Telco |

*Table 2.* GA Parameters

## 6. Results

The graph compares the performance of three crossover methods—single-point, uniform, and two-point—in a genetic algorithm over 20 generations. Single-point crossover shows steady improvement, ending up close to the others. The uniform crossover starts and ends with the highest fitness values, demonstrating robust improvement, especially in early generations. Two-point crossover maintains a middle ground between single-point and uniform methods.

Figure 3 depicts Error bars indicating decreasing variability as generations progress, implying convergence towards stable solutions. On average, the Uniform method produced the best solution, boasting an average score of approximately -6717.541. In comparison, the Single Point method achieved an average score of -6730.157, while the Two Point method attained an average of -6715.126 as shown in table 3. This data underscores the superiority of the Uniform method in yielding higher fitness values across the generations studied.

| Methods | Best Fitness |
|---|---|
| Single-Point Crossover | -6730.157 |
| Uniform Crossover | -6717.541 |
| Two-Point Crossover | -6715.126 |

*Table 3.* Candidate Method Result

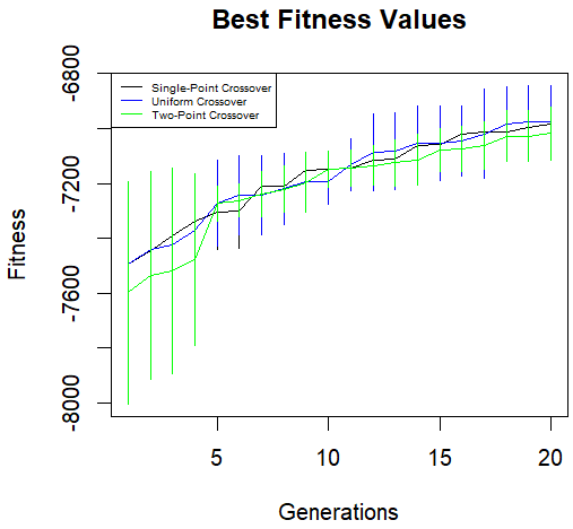


*Figure 4.* Best Fitness Results for Each Candidate Method over 20 Iterations

## 7. Discussion

Alongside the recognized efficacy of the Uniform Crossover method, logistic regression, a conventional yet potent technique, could serve as a benchmark for contrasting with more intricate models like artificial neural networks (ANN). Expanding beyond Uniform to incorporate various crossover methods such as Single Point or Multi-Point may offer additional insights into feature selection strategies within genetic algorithms. By amalgamating the interpretability of logistic regression, the predictive capabilities

of ANN, and the feature selection prowess of genetic algorithms through diverse crossover methods, the potential arises for more robust churn prediction models. These models could substantially enhance customer retention strategies in practical scenarios.

While both single and uniform crossovers exhibited strong performance, there exists an expectation for improved results with two-point crossover. Such a multifaceted approach holds promise for optimizing churn prediction models, thereby empowering telecom companies with better decision-making tools to mitigate customer attrition and bolster business performance.

## 8. Conclusion and Future Work

In conclusion, the evaluation of predictive models' performance using error bar plots underscored the superiority of the Uniform Crossover method over alternative techniques, affirming its adeptness in selecting pertinent features for customer churn prediction. The optimal solution unveiled in the chromosome sequence offered invaluable insights into the pivotal features influencing churn prediction outcomes. Moving forward, prospective research avenues could explore alternative feature selection methodologies, delve into ensemble techniques, and integrate external factors to bolster model generalizability and transparency. Such endeavors hold the potential to streamline decision-making processes for telecom companies by furnishing them with more robust and interpretable predictive models.

Furthermore, future investigations might delve deeper into comprehending the underlying mechanisms behind these unexpected findings, facilitating the refinement of hypotheses and experimental methodologies to enhance outcomes in subsequent research endeavors, thereby advancing the field of telecom customer churn prediction.

## References


Abualigah, L. and Dulaimi, A. J. A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm. *Cluster Computing*, 24: 2161–2176, 2021.

Agrawal, S., Das, A., Gaikwad, A., and Dhage, S. Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In *2018 International conference on smart computing and electronic enterprise (ICSCEE)*, pp. 1–6. IEEE, 2018.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., and Anwar, S. Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, 94:290–301, 2019.

Babatunde, O. H., Armstrong, L., Leng, J., and Diepeveen, D. A genetic algorithm-based feature selection. 2014.

Fridrich, M. Hyperparameter optimization of artificial neural network in customer churn prediction using genetic algorithm. *Trends Economics and Management*, 11(28): 9–21, 2017.

Fujo, S. W., Subramanian, S., Khder, M. A., et al. Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, 11(1):24, 2022.

Jain, H., Khunteta, A., and Srivastava, S. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 167:101–112, 2020.

Liu, A., Ghosh, J., and Martin, C. Generative oversampling for mining imbalanced datasets. *DMIN*, 7:66–72, 2007.

Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., and Azmi, A. Customer churn prediction in telecommunication industry using machine learning classifiers. In *Proceedings of the 3rd international conference on vision, image and signal processing*, pp. 1–7, 2019.

Momin, S., Bohra, T., and Raut, P. Prediction of customer churn using machine learning. In *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018*, pp. 203–212. Springer, 2020.

Pamina, J., Raja, B., SathyaBama, S., Sruthi, M., VJ, A., et al. An effective classifier for predicting churn in telecommunication. *Jour of Adv Research in Dynamical & Control Systems*, 11, 2019.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., and Snoeck, M. Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, 40:116–130, 2018. ISSN 2210-6502. doi: https://doi.org/10.1016/j.swevo.2017.10.010. URL https://www.sciencedirect.com/science/article/pii/S2210650216301754.

Xue, Y., Zhu, H., Liang, J., and Słowik, A. Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification. *Knowledge-Based Systems*, 227:107218, 2021.

Zaharie, D. Influence of crossover on the behavior of differential evolution algorithms. *Applied soft computing*, 9(3): 1126–1138, 2009.