YILDIZ TECHNICAL UNIVERSITY

FACULTY OF CHEMISTRY AND METALLURGY

DEPARTMENT OF MATHEMATICAL ENGINEERING

MULTIDISCIPLINARY DESIGN PROJECT

APPLICATIONS OF INTERPOLATION IN MACHINE LEARNING

Thesis Supervisor: Assoc. Prof. Dr. B. Ali İBRAHİMOĞLU

21058025 CANER ERENLER

Istanbul, 2026

TABLE OF CONTENTS Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under Curve
CART	Classification and Regression Trees
CatBoost	Categorical Boosting
IG	Information Gain
IQR	Interquartile Range
KNN	K-Nearest Neighbor
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
MCAR	Missing Completely at Random
MNAR	Missing Not At Random
NaN	Not a Number
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive Explanations
SSE	Sum of Squared Errors
SVM	Support Vector Machines
XGBoost	eXtreme Gradient Boosting
F1	F1 Score
MAE	Mean Absolute Error
MSE	Mean Squared Error

PREFACE

I would like to thank my Thesis Supervisor, Assoc. Prof. Dr. B. Ali İBRAHİMOĞLU, for his guidance and direction throughout this research. I also extend my deepest gratitude to my dearest father, Oktay Erenler. I sincerely thank my mother, Serap Erenler, and my sister, Cansu Erenler, for their unwavering support and patience. I extend my appreciation to my friends Yasemin Kasapoglu and Bengu Serin for keeping my motivation alive with their encouragement, and to Idil Merve Aksahin for her support and contributions during this period.

ABSTRACT

Data quality is critical for predictive modelling, but missing values invariably compromise the integrity of the model. Traditional imputation methods often distort data distributions or result in the loss of useful information. This thesis examines a novel 'Pseudo Time Series' approach that leverages internal variable correlations to repurpose interpolation techniques, typically reserved for temporal data, for cross-sectional datasets. The main methodology involves ordering the data according to a highly correlated auxiliary variable, thereby enabling deterministic imputation via Linear, Polynomial, and Cubic Spline interpolation.

Using the Telco Customer Churn dataset, a 10% Missing Completely at Random (MCAR) model was applied to the TotalCharges variable. Thanks to a strong correlation with Tenure ($r \approx 0.825$), the dataset was ordered, creating a sequential series close to the TotalCharges = f(Tenure) approach. The accuracy of interpolation-based imputation was compared systematically with standard methods (Mean Imputation, KNN, Least Squares) in eight machine learning models and evaluated using 10-fold cross-validation with ROC_AUC as the main metric.

Key findings show that Polynomial Interpolation performed best with Logistic Regression (ROC_AUC: 0.8463), even outperforming the model trained on the original, complete dataset. This points to an additional smoothing effect that reduces noise and increases generalisation. Mean Imputation performed consistently low as a result of artificial variance suppression, while Spline and KNN methods maintained the data structure. A notable model-method compatibility was observed: global interpolators performed well with global models (e.g., Polynomial with Logistic Regression), while local imputations performed well with local learners (e.g., KNN with LightGBM). Importantly, Cubic Spline interpolation provides a scalable and efficient trade-off, delivering competitive accuracy at a computational cost considerably lower than KNN.

These findings affirm that interpolation is a practical and effective imputation strategy for sequential cross-sectional data, performing better than existing traditional approaches and preserving critical data relationships. This research provides a practical, correlation-focused approach to improving data quality in machine learning processes, which is of particular importance for customer analytics in sectors such as telecommunications, finance, and healthcare.

Keywords: Missing Data, Imputation, Interpolation, Cross-sectional Data, Pseudo Time Series, Customer Churn, Machine Learning, Data Preprocessing

ÖZET

Veri kalitesi, tahmin modellemesi için çok önemlidir, ancak eksik değerler modelin bütünlüğünü her zaman bozar. Geleneksel imputasyon yöntemleri genellikle veri dağılımlarını çarpıtır veya değerli bilgilerin kaybolmasına neden olur. Bu tez, içsel değişken korelasyonlarından yararlanarak, genellikle zamansal veriler için ayrılmış olan enterpolasyon tekniklerini kesitsel veri kümeleri için yeniden kullanan yeni bir "Sözde Zaman Serisi" yaklaşımını incelemektedir. Ana metodoloji, yüksek korelasyonlu bir yardımcı değişkene göre verileri sıralamayı içerir ve böylece Doğrusal, Polinom ve Kübik Spline interpolasyonu yoluyla deterministik imputasyonu mümkün kılar.

Telco Customer Churn veri seti kullanılarak, TotalCharges değişkenine %10 Tamamen Rastgele Eksik (MCAR) modeli uygulanmıştır. Tenure ile güçlü bir korelasyon ($r \approx 0,825$) sayesinde veri seti sıralanmış ve TotalCharges = f(Tenure) yaklaşımına yakın bir sıralı dizi oluşturulmuştur. Interpolasyon tabanlı imputasyonun performansı, sekiz makine öğrenimi modelinde standart yöntemlerle (Ortalama Imputasyon, KNN, En Küçük Kareler) sistematik olarak karşılaştırıldı ve birincil metrik olarak ROC_AUC kullanılarak 10 kat çapraz doğrulama ile değerlendirildi.

Kilit sonuçlar, polinom enterpolasyonu lojistik regresyon ile en yüksek performansı (ROC_AUC: 0.8463) elde ettiğini ve hatta orijinal, tam veri seti üzerinde eğitilmiş modelin performansını aştığını ortaya koymaktadır. Bu, gürültüyü azaltan ve genellemeyi artıran ilave bir yumuşatma etkisi olduğunu işaret etmektedir. ortalama atama, yapay varyans baskılaması sonucu tutarlı bir şekilde düşük performans gösterirken, Spline ve KNN yöntemleri veri yapısını korumuştur. Dikkate değer bir model-yöntem uyumu ortaya çıktığı görülmüştür: global enterpolatörler global modellerle (ör. Lojistik Regresyon ile Polinom) üstün performans gösterirken, yerel imputasyonlar yerel öğrenenlerle (ör. LightGBM ile KNN) iyi bir uyum sağlamıştır. Önemli olarak, Kübik Spline enterpolasyonu ölçeklenebilir ve verimli bir uzlaşma sağlarken, KNN'ye göre önemli ölçüde daha düşük hesaplama maliyetiyle rekabetçi bir doğruluk sağlamıştır.

Bulgular, enterpolasyonun sıralı kesitsel veriler için uygulanabilir ve etkili bir imputasyon stratejisi olduğunu, mevcut geleneksel yaklaşımlardan daha iyi performans gösterdiğini ve kritik veri ilişkilerini muhafaza ettiğini teyit etmektedir. Bu çalışma, telekomünikasyon, finans ve sağlık gibi sektörlerdeki müşteri analitiği için özel öneme sahip, makine öğrenimi süreçlerinde veri kalitesini artırmak için pratik, korelasyon odaklı bir çerçeve sunmaktadır.

Anahtar Kelimeler: Eksik Veriler, İmpütasyon, Enterpolasyon, Kesitsel Veriler, Sözde Zaman Serileri, Müşteri Kaybı, Makine Öğrenimi, Veri Ön İşleme

1. INTRODUCTION

Nowadays, organisations must analyse large amounts of data in order to make strategic decisions. Especially in areas where customer acquisition costs are high, such as the telecommunications sector, customer retention is critical importance [1]. In this context, customer churn prediction stands out as one of the most common and strategic application areas of machine learning methods [2]. However, customer databases frequently contain missing data due to systemic errors, disruptions in data collection processes, privacy preferences, or integration problems between different systems [3]. These shortcomings directly jeopardise the developed prediction models' accuracy and overall relevance[4].
Machine learning algorithms aim to learn a prediction function based on observed data. This is mathematically expressed as a mapping problem defined mathematically as

$$f: X \rightarrow Y$$

Here, $X \in \mathbb{R}^{n \times p}$represents the input matrix consisting of n observations and p features, while Y represents the target variable to be predicted.
Under optimal conditions, it is expected that all elements of the input matrix X are fully observed. However, in real-world applications, this assumption often cannot be met, and some observations contain missing values. This situation is mathematically expressed as

$$x_{ij} = \emptyset$$

this is referred to as the missing data problem in the literature[5]. The existence of missing data has a direct impact on the training process of machine learning models and limits the generalisability of the learned function [6].
One of the most common approaches to the problem of missing data is to completely remove rows containing missing observations from the dataset. Mathematically, this method is

$$X_{new} = \{ x_i \mid \forall j,\, x_{ij} \neq \emptyset \}$$

defined as such and referred to in the literature as listwise deletion [7]. However, this approach reduces the number of effective observations by narrowing the sample space, which in turn leads to a decrease in statistical power [8]. Particularly in datasets with high rates of missingness, this method severely limits the model's learning capacity[5].

Another approach frequently used to solve the missing data problem is to fill in the missing values with the expected value of the relevant variable. Known as mean imputation, this method:

$$\hat{x}_{ij} = \mathbb{E}[X_j]$$

This method is defined as follows. Although it's easy to use and doesn't require much computing power, this approach has significant effects on the data distribution [5]. Replacing missing values with the mean causes the data to cluster at a single point, which artificially reduces the variable's variance [6]. This effect is stronger when more data is missing, and if almost all the data is missing, the variance is expected to approach zero.

$$\sigma^2 \to 0$$

It is further highlighted in [5] that this circumstance introduces bias into the dataset, thereby diminishing model efficacy, especially within variance-sensitive machine learning algorithms.

The central concern of this thesis is to address the challenge of missing data not solely through statistical imputation, but by maintaining the mathematical and structural interdependencies inherent in the data. More specifically, the investigation centres on whether the physical and logical relationships among variables exemplified by the correlation between total payments and subscription duration (Tenure) can be effectively leveraged in the estimation of missing data, particularly within cross-sectional customer datasets that lack time series characteristics.
The core question is whether powerful interpolation methods, like Polynomial and Cubic Spline techniques, which are often used in time series analysis, can be applied to cross-sectional data. This requires a suitable way to order the data [9].

Within this scope, the aim of the study is to examine the effects of polynomial interpolation, cubic spline interpolation, and regression-based imputation methods on missing data estimation by ranking a non-ordered customer dataset based on a highly correlated variable (Tenure), and to systematically analyse the contribution of these methods to model success in classification problems such as customer churn prediction. Thus, the goal is to preserve the statistical properties of the dataset while improving the classification performance of machine learning models [5].

This study aims to examine the mathematical and statistical effects of missing data imputation methods on cross-sectional customer data and tests the following hypotheses:

H1: Interpolation methods are mathematically applicable and produce meaningful results on cross-sectional data sets ranked using highly correlated variables[10].

H2: Interpolation-based imputation methods improve the performance of classification models compared to the traditional Mean Imputation method [3].

H3: The success of the imputation method depends on the mathematical structure and learning mechanism of the machine learning model used [11].

The ‘Pseudo-Time Series’ approach used in this study does not represent a time-dependent dynamic process modelling in the classical sense. The aim here is to use a variable (Tenure) with high correlation and meaningful physical interpretation as an

independent ordinal axis in a cross-sectional data set. In this context, the TotalCharges variable:

$$\text{TotalCharges} \approx f(\text{Tenure}) + \varepsilon$$

Modelled in this form; missing values have been estimated based on a deterministic approach to this functional relationship rather than temporal dependence. Therefore, the applicability of interpolation methods depends not on time series assumptions but on the presence of a well-defined, relatively smooth and uniform functional structure on the independent variable [10].

2. THEORETICAL BACKGROUND AND MATHEMATICAL FOUNDATIONS

This section details the mathematical foundations of the missing data simulation mechanism used in the study, the transformation techniques applied for data manipulation, and the deterministic and stochastic algorithms used for missing value estimation. The parameter selections for the methods are justified based on the implementation details in the Python Scikit-learn and Pandas libraries.

2.1. Missing Data Mechanism and Simulation

The fundamental assumption in missing data analysis is modelling the mechanism of missingness. According to Little and Rubin [5] theory, when the data matrix Y is divided into the observed part Y_{obs} and the missing part Y_{mis}, the probability distribution of the missingness indicator matrix M is defined as follows:

$$P(M \mid Y, \phi)$$

In this study, the Missing Completely at Random (MCAR) mechanism was used to measure the pure mathematical performance of algorithms and prevent bias. Under the MCAR assumption, the probability of missingness is independent of the data values:

$$P(M \mid Y, \phi) = P(M \mid \phi)$$

In the code implementation, the constant random_state = 123 was used to ensure experimental reproducibility. Using an approach similar to Monte Carlo simulation, a random 10% subset was selected from the sample space of the *TotalCharges* variable and marked as np.nan.

2.2 Pseudo-Time Series Transformation

Interpolation methods require a sequential structure on the definition set ($x_1 < x_2 < \cdots < x_n$). Since the data set used is cross-sectional in nature, there is no natural time axis. However, there is a strong positive linear relationship between the TotalCharges variable and the Tenure variable.

To quantitatively validate this relationship, the Pearson correlation coefficient was calculated and obtained as $r = 0.825$. This value indicates a strong positive linear relationship between the two variables. The Pearson correlation coefficient was calculated using the .corr() function of the Pandas library, and the calculation is based on the default Pearson definition.

Taking advantage of this structural relationship, the data set was transformed into a monotonically increasing sequence according to the Tenure variable:

$$S = \{ (x_i, y_i) \in D \mid x_i \leq x_{i+1}, \forall i \}$$

This transformation mathematically reduced the static data set to a continuous function approximation in the form $y = f(x)$ and ensured the applicability of interpolation techniques.

2.3 Applied Interpolation and Imputation Methods

In this study, the problem of missing data is addressed under the heading of imputation (value assignment), in line with the general classification in the literature. As is known, imputation methods may include statistical, stochastic or deterministic approaches. In this context, the interpolation methods used in the study (Linear, Polynomial and Cubic Spline) have been evaluated as a special subclass of deterministic imputation methods because they derive missing values using the mathematical function relationship between neighbouring observations in the sorted data set. Therefore, throughout the thesis, the term "interpolation" is used to refer to the mathematical operator used to fill in missing data, while the term "imputation" is used to refer to the general solution process of the problem, within a hierarchical relationship.

The methods presented in this section approach the problem of missing data from different mathematical perspectives. While mean imputation and KNN methods are statistical (distribution-based) or topological (distance-based) imputation strategies, Linear, Polynomial, and Cubic Spline methods are deterministic-functional imputation approaches based on modelling the data as a continuous function under an ordered structure. In line with this structural distinction, all methods were evaluated within a common framework in the comparative analysis.

2.3.1 Simple Mean Imputation

In this method, missing values are filled in with the expected value of the series:

$$\hat{x}_i = \mu = \frac{1}{n} \sum_{j=1}^{n} x_j$$

As this approach tends to reduce variance, it has been used as the baseline method for comparison purposes in the study.

2.3.2 Linear Spline Interpolation

Linear spline interpolation is a piecewise function approximation defined on a known set of ordered data points, which is continuous but not differentiable.

Considering a set consisting of the $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ with the ordering $a = x_0 < x_1 < \cdots < x_n = b$, the linear spline function $S(x)$ is defined as a first-degree polynomial (a straight line) on each subinterval $[x_i, x_{i+1}]$ subinterval.

For each subinterval $i = 0,1, \dots, n-1$ in $x \in [x_i, x_{i+1}]$, the interpolation function is expressed by the following formula:

$$S_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i)$$

In this equation:

- $S_i(x)$: the interpolation value in the i-th interval,
- y_i: the function value at the point x_i,
- $\frac{y_{i+1}-y_i}{x_{i+1}-x_i}$: represents the gradient of the line in the relevant interval.

The global function $S(x)$ obtained by this method is continuous at the knot points $S(x_i^-) = S(x_i^+)$, but its first derivative is discontinuous at these points due to the changes in slope ($S'(x_i^-) \neq S'(x_i^+)$). This situation creates an angular structure in the graph of the function.

2.3.3 Polynomial Interpolation

A polynomial of degree n is used to model the general trend of the data. Lagrange form:

$$P(x) = \sum_{j=0}^{n} y_j \, L_j(x)$$

Due to the risk of the Runge phenomenon in high-order polynomials, a second-order polynomial (order = 2) was preferred in the study.

2.3.4 Cubic Spline Interpolation

In this study, the Cubic Spline method used to complete missing values in the data set creates a smooth curve of class C^2 (continuous up to the second derivative) by defining third-degree polynomials between data points.

When $n+1$ data points $(\mathrm{x}_0, \mathrm{y}_0), \ldots, (\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})$ are given, the function $S_{i(x)}$ defined for $[x_i, x_{\{i+1\}}]$ in each sub-interval $i = 0, 1, \ldots, n-1$ is expressed in the following general form:

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

Since there are four unknown coefficients a_i, b_i, c_i, d_i for each interval in this system, a total of $4n$ unknowns must be solved. To arrive at a unique solution, the system is based on the following derivative equations and boundary conditions.

The interpolation conditions of the spline function must satisfy the actual values in the data set at the knot points. This condition, when $x = x_i$, indicates that the coefficient d_i is directly equal to the value y_i:

$$S_i(x_i) = y_i \text{ and } S_i(x_{i+1}) = y_{i+1}$$

The conditions of continuity and differentiability ensure that the piecewise nature of the curve is not apparent and that a smooth transition is achieved. At the internal nodal points, the function $x_1, \ldots, x_{n-1}$ itself, its first derivative, and its second derivative must be continuous:

- Function Continuity (C^0):

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1})$$

- First Derivative Continuity (Slope Equality, C^1):

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$$

- Second Derivative Continuity (Curvature Equality, C^2):

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$$

Boundary conditions must be defined at the endpoints to ensure the solubility of the problem (closure of the degree of freedom). In this study, the 'Natural Spline' condition

was adopted to preserve data variance and prevent artificial oscillations at the endpoints. This condition assumes that the second derivative (curvature) is zero at the endpoints:

$$S_0''(x_0) = 0 \text{ and } S_{n-1}''(x_n) = 0$$

When the matrix formulation is substituted into equation $S_i(x)$ under the above conditions, the unknown coefficients (usually second derivative moments in terms of M_i) are linked together. As a result of this process, the problem is transformed into a tridiagonal (three-banded) linear equation system that can be efficiently solved using numerical analysis:

$$Ax = b$$

Here, matrix A has a tridiagonal structure:

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

The solution to this linear system provides the optimal coefficients that minimise the following integral:

$$\min \int_{x_0}^{x_n} [S''(x)]^2 dx$$

This minimisation is the mathematical proof of why the Cubic Spline method smooths out sudden changes (noise) in the data set while preserving the main trend.

2.3.5 K-Nearest Neighbour (KNN Imputation)

In this method, the missing observation is estimated using the average of the k nearest neighbours in the feature space. The distance metric used is the Euclidean norm:

$$d(x_q, x_i) = \| x_q - x_i \|_2$$

In this study, $k = 5$ has been selected.

2.3.6 Least Squares Regression

The Least Squares method is a deterministic approach to modelling the relationship between data. The fundamental objective of the method is to minimise the sum of the squares of the differences between the observed actual values (y_i) and the values predicted by the model $f(x_i)$.

The total error function (E) is defined as follows for n data points:

$$E = \sum_{i=0}^{n} [y_i - f(x_i)]^2$$

In this study, three different model structures were examined to minimise the error function (Target: $\min E$):

Firstly, in linear models that situation where there is a linear relationship between variables. Secondly, polynomial models that Situations where curvilinear relationships are expressed by nth degree polynomials. Thirdly, data linearisation that Situations where non-linear structures, such as exponential or logarithmic, are transformed into linear form with the help of transformations.

All three approaches can be expressed in general matrix form as $Y = X\beta + \varepsilon$ Here, the objective function takes the following form in matrix notation:

$$J(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Taking the derivative of this function with respect to the parameter β and setting it equal to zero yields the general solution (Normal Equations) for the parameter vector that minimises the error ($\hat{\beta}$):

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

2.4 Model Validation Framework

The success of the developed imputation strategies was evaluated using the 10-fold cross-validation method. This approach was chosen to ensure bias–variance balance and to obtain statistically reliable results.

3. MATERIAL AND METHODS

This section presents the data set topology forming the experimental framework of the research, controlled missing data simulation, the mathematical model of the proposed Pseudo-Time Series transformation, and the technical details of the six different missing data imputation algorithms applied. The implementation process was carried out using the Python programming language; Pandas and Numpy were used for data manipulation, while the Scikit-learn, LightGBM, CatBoost, and XGBoost libraries were used for machine learning algorithms.

3.1 Topological Properties of the Data Set and Preprocessing

The data set used in the study is the 'Telco Customer Churn' data set, which contains anonymised customer records from a company operating in the telecommunications sector. The data set is defined in a matrix space consisting of **7043** observation vectors and **21** feature dimensions:

$$X \in \mathbb{R}^{7043 \times 21}$$

The feature space $\mathcal{F}$ in the dataset consists of the union of three subspaces ($\mathcal{F} = \mathcal{F}_{demo} \cup \mathcal{F}_{serv} \cup \mathcal{F}_{acc}$):

Demographic Variables ($\mathcal{F}_{demo}$): Gender, SeniorCitizen, Partner, Dependents.

Service Variables ($\mathcal{F}_{serv}$): PhoneService, MultipleLines, InternetService, OnlineSecurity, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.

Account and Payment Variables ($\mathcal{F}_{acc}$): Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges, Tenure.

The target variable is the Churn column, which indicates whether the customer has churned or not, and is a binary vector following a Bernoulli distribution defined on the $\{0, 1\}$ set. Within the load_and_prep_data function in the code block, the Churn variable has been mapped from the 'Yes'/'No' format to the $\{1, 0\}$ format.

3.1.1 Data Type Conversions and Initial Cleaning

In the raw dataset, the TotalCharges variable should be a numeric vector, but it is stored as an object type due to the empty character strings (‘ ’) it contains. Prior to analysis, this variable was converted to numerical form using the $f: S \to \mathbb{R}$ conversion (pd.to_numeric), and characters that could not be converted (errors=“coerce”) were marked as NaN (Not a Number). These very few natural deficiencies at the outset were filled using the median imputation method to avoid compromising the consistency of the simulation, and the data set was transformed into X_{full}.

3.2 Experimental Simulation Design: MCAR Mechanism

To measure the performance of missing data imputation methods in an unbiased manner, the Controlled Missingness Simulation method, which is widely accepted in the literature, was applied. In this study, the Missing Completely at Random (MCAR) mechanism, where missingness is independent of the data itself and other variables, was adopted.

The mathematical steps of the simulation are as follows:

Reference Set: The complete data set X_{full} is taken as a reference.

Deterministic Initialisation: To ensure the reproducibility of the experiment, the stochastic process was made deterministic using the np.random.seed(123) constant.

Index Selection: Random indices (missing_indices) are selected from the sample space, which is a subset of the index set $I = \{1,2, \dots, N\}$ and satisfies the cardinality condition $| I_{miss} | = \lfloor 0.10 \times N \rfloor$.

Masking: The TotalCharges values at the selected indices are deleted to create the artificial missing X_{miss} matrix:

$$X_{miss}(i,\text{TotalCharges}) = \begin{cases} \text{NaN}, & i \in I_{miss} \\ X_{full}(i,\text{TotalCharges}), & \text{otherwise} \end{cases}$$

The dataset obtained as a result of this process forms the raw material from which the different datasets detailed below will be derived.

The missingness rate has been selected as 10%; this rate represents the moderate level of data loss frequently encountered in real-world customer databases and provides a balanced scenario for observing the behaviour of imputation methods.

3.3. Proposed Method: Correlation-Based Ranking and Pseudo-Time Series

The most original methodological approach of this thesis is modelling cross-sectional data, which does not possess time series characteristics, as if it were a time series using the physical and statistical correlations between them.

Interpolation methods (Linear, Polynomial, Spline) require a monotonically increasing ordered structure $(x_1 < x_2 < \cdots < x_n)$ in the domain. Our data set does not have a natural time axis; however, there is a mathematically strong positive linear relationship between TotalCharges (Y) and Tenure (Customer Duration, (X)).

The strength of this relationship was calculated using the .corr() function in the Pandas library and expressed using the Pearson Correlation Coefficient formula as follows:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \approx 0.825$$

Here, $\bar{x}$ and $\bar{y}$ are the sample means of the Tenure and TotalCharges variables, respectively.

This high correlation value $(r \approx 0.825)$, statistically justifies that when the data is sorted according to the Tenure variable, the TotalCharges variable has a distinct trend component that can be modelled as $Y \approx f(X) + \varepsilon$, rather than a purely random distribution.

The transformation applied in the code block (df_sorted) is as follows:

$$D_{sorted} = \text{sort\_values}(D_{miss}, \text{by} = \text{Tenure})$$

After this process, the data set is sorted according to subscription duration $(from\ t = 0 \text{ to } t = 72)$rather than by index, and acquires a topology suitable for forward interpolation algorithms.

3.4. Applied Imputation Algorithms and Mathematical Models

In this study, a total of six different mathematical approaches were applied to the created A and B data sets, resulting in a total of seven different variations (original data set + 6 imputation methods = 7 data set variations).

In this study, missing data simulation was applied only to the TotalCharges variable. The main reason for this is that this variable has a high correlation structure with variables such as Tenure and MonthlyCharges and also plays a critical role in customer churn prediction. Multivariate missingness scenarios were excluded from the scope of this study in order to isolate the behaviour of the method and limit experimental complexity.

3.4.1. Dataset 2: Simple Mean Imputation

The second data set (df_mean) was configured using the simple mean imputation method, which is the most fundamental statistical approach in the literature and was positioned as the 'baseline' for comparing the success of other methods in this study. In the code implementation, missing observations in the TotalCharges variable were filled with a single fixed value calculated as the overall arithmetic mean of the series using the fillna() function. Although this method has a tendency to artificially reduce the variance of the data and weaken the correlation between variables (centring), it was included in the analysis in order to concretely measure the performance increase provided by the advanced interpolation and regression-based methods proposed in the study and to establish a lower bound.

3.4.2. Dataset 3: Linear Spline Interpolation

When creating the third data set (df_linear), missing data was completed using the Linear Spline Interpolation method. Before applying this process in the code, the entire dataset was sorted according to the tenure variable in order to accurately model the temporal relationship and growth trend between the data points. Then, using the interpolate(method="linear")function, each missing value was filled in with its equivalent on a virtual line drawn between the immediately preceding and following known points. This method creates a C^0 continuous (smooth but not differentiable) structure by following local changes in the data, but it is limited in modelling the overall curvilinear behaviour of the data. In the final stage, the missing values at the beginning and end of the data set, where it is not possible to draw a line between two points, were addressed by adding forward (ffill) and backward (bfill) filling functions to the code block, thus completing the process.

3.4.3. Dataset 4: Piecewise Cubic Spline

This method forms the basis of the thesis's fundamental hypothesis and has been applied to the df_spline dataset. Unlike global polynomial fitting procedures, this approach models the dataset using piecewise polynomials defined separately for each data range, rather than a single function.

In the Python implementation, the parameters method="spline" and order=3 were specifically selected for this study. The choice of order=3 is critical for this project because it is the lowest degree required to ensure mathematical C^2 continuity 1. This specific configuration allows the model to preserve the natural variance of the Telco Churn dataset and smooth out oscillations without being subject to the rigidity of lower-order methods or the oscillation problems caused by higher-order methods. It has enabled the model to preserve the natural variance of the Telco Churn dataset and smooth out abrupt changes.

3.4.4. Dataset 5: Polynomial Interpolation

When configuring the fifth data set, the 'Polynomial Interpolation' strategy (df_poly) was applied to fill in the missing TotalCharges values, going beyond linear approaches. In this process, the data set was first sorted from smallest to largest according to the tenure variable, which represents the subscription period, to ensure that the interpolation could produce mathematically meaningful results. As model parameters, method="polynomial" and order=2 (quadratic) were selected to capture the curvilinear increase (convex structure) inherent in the data. To prevent the severe oscillation problem created by high-degree polynomials, particularly at extreme values, known in the literature as the 'Runge phenomenon,' the polynomial degree was deliberately limited to 2, and the model $P(x) = a_2x^2 + a_1x + a_0$ was adopted. Finally, to fill the gaps at the boundary points (end values) where the mathematical model cannot perform calculations, backward (bfill) and forward (ffill) completion steps were integrated into the algorithm to ensure data integrity.

3.4.5. Dataset 6: K-Nearest Neighbour (KNN Imputation)

When creating the sixth dataset (df_knn), the K-Nearest Neighbour (KNN) algorithm was used, which does not make a parametric assumption about the distribution of the data and is based on the topological similarity between observations. In the code implementation, the Scikit-learn library's KNNImputer module was used, and the number of neighbours parameter determining the model's accuracy was fixed at n_neighbours=5. This method bases itself on the Euclidean distance (L2 Norm) in multidimensional space rather than looking at a single variable or temporal sequence when filling in missing data. During implementation, to ensure consistent and noise-free distance calculations, the algorithm was run using only the numerical variables directly related to the missing data (tenure, MonthlyCharges, and TotalCharges) rather than the entire dataset. Thus, missing

values were completed by taking the weighted average of the five closest neighbours with similar payment habits and subscription durations.

3.4.6. Dataset 7: Least Squares Regression

When creating the seventh dataset (df_ls), the Least Squares Regression method, a deterministic approach that takes advantage of the strong linear relationship between variables, was applied. On the code side, considering the principle that the TotalChargesvariable is mathematically directly related to the tenure and MonthlyCharges variables, the imputation process was designed as a multivariate regression problem. During the application phase, the data set was divided into two logical parts: rows with TotalCharges data were used as the 'training set' for the model to learn the coefficients, while rows with missing values were used as the 'prediction set'. The LinearRegression algorithm from the scikit-learn library created an equation to predict the TotalCharges output using the tenure and MonthlyCharges inputs on the training set and used this equation to fill in the missing data. This method is one of the imputation methods with the highest consistency between variables, as it does not involve randomness and preserves the physical structure of the data (covariance matrix).

3.5. Feature Engineering Pipeline

Each of the seven different datasets, with imputation completed, was passed through a standard pre-processing pipeline using the feature_engineering_pipeline function before being fed into machine learning models.

3.5.1. Outlier Handling

The IQR (Interquartile Range) method was used to reduce statistical noise in the dataset. Lower $(Q_1 - 1.5 \times IQR)$ and upper $(Q_3 + 1.5 \times IQR)$ threshold values were calculated for numerical variables, and values outside these limits were capped to the limits.

3.5.2. Feature Extraction

To indirectly test the success of the imputation methods, new variables were created derived from the imputed TotalCharges variable. For example:

$$\text{NEW\_AVG\_Charges} = \frac{\text{TotalCharges}}{\text{Tenure} + 1}$$

$$\text{NEW\_Increase} = \frac{\text{NEW\_AVG\_Charges}}{\text{MonthlyCharges}}$$

If the imputation method fails (e.g., Mean Imputation), the derived NEW_AVG_Charges variable will also be erroneous and have low information gain, as the variance of the TotalCharges variable will be low. These variables were used as indicators measuring imputation quality in the model's importance ranking (Feature Importance).

3.5.3. Variable Encoding

Categorical variables were converted into numerical matrices to enable mathematical operations by the algorithms:

• Label Encoding: Variables with cardinality 2 (e.g., Partner, Dependents) were mapped to the $\{0, 1\}$ scale.

• One-Hot Encoding: Nominal variables (e.g., InternetService) were encoded by creating $k - 1$ vectors for each class using the pd.get_dummies(drop_first=True) function to reduce the risk of 'dummy variable trap' and multicollinearity.

In this study, in order to isolate the effect of missing data imputation methods on model performance, all created datasets were processed through the same and fixed feature engineering pipeline. This approach ensured that the observed performance differences were attributed to imputation strategies rather than feature extraction or transformation.

3.6. Modelling and Validation Strategy

In this study, missing data imputation procedures were applied to the entire dataset prior to the cross-validation process. This approach is theoretically considered in the literature to be an application that carries the risk of *data leakage*. However, the main objective of this study is to comparatively examine the relative effects of different imputation methods rather than to maximise absolute model performance.

Therefore, all imputation methods and all machine learning models were evaluated under the same experimental conditions; there was no difference in access to information that would give any method an advantage. Consequently, the findings obtained maintain their validity in terms of the comparative performance analysis of imputation methods rather than absolute generalisation performance.

3.6.1. Model-Agnostic Approach

In this study, eight algorithms based on different mathematical foundations were tested to demonstrate that the success of the imputation method is independent of the classifier used:

Linear Models: Logistic Regression (LR).

Distance-Based: K-Nearest Neighbours (KNN).

Knowledge-Based (Trees): Decision Tree (CART), Random Forest (RF).

Geometric: Support Vector Machines (SVM).

Boosting (Gradient-Based): XGBoost, LightGBM, CatBoost.

3.6.2. 10-Fold Cross Validation

To ensure the statistical reliability of the results, the cv=10 parameter was used in the cross_validate function. The data set was divided into 10 separate parts $\{D_1, D_2, \dots, D_{10}\}$, with 9 parts used as the training set and 1 part used as the test set in each iteration.

As a performance metric, the ROC_AUC (Receiver Operating Characteristic - Area Under Curve) score, which best reflects accuracy in churn datasets with class imbalance, was used.

Since ROC_AUC is a threshold-independent criterion, it provides more stable and reliable results than the Accuracy metric alone in class imbalance. The aim is to determine which imputation method (Spline, Mean, etc.) maximises the area under the ROC curve.

3.6.3. Imputation Timing and Data Leakage Assessment:

In this study, missing data imputation procedures were applied to the entire dataset prior to the cross-validation process. This approach may theoretically carry a risk of data leakage, as noted in the literature. However, the primary objective of the study is to comparatively examine the relative performance of different imputation methods rather than maximising absolute model performance. Therefore, all methods and models were evaluated under the same level of information and the same experimental conditions. Consequently, the results obtained maintain their validity in terms of the comparative effectiveness of imputation methods rather than absolute generalisation performance.

4 RESULTS AND ANALYSIS

This section presents the performance results of eight different machine learning algorithms trained on seven different data set variations (Original, Mean, Linear, Spline, Polynomial, KNN, Least Squares) detailed in previous sections. The analyses are based on the average **Accuracy**, **F1-Score**, and **ROC_AUC** metrics obtained as a result of 10-fold cross-validation.

The analysis process was structured around three main axes:

Numerical Performance Analysis: Which imputation method achieved the global maximum success with which algorithm?

Feature Importance Analysis: What is the weight of the variables in the model's decision mechanism?

Explainability (SHAP) Analysis: Does the relationship established by the model correspond to physical reality?

4.1. Comparative Analysis of Model Performance

The results of the 56 different experimental scenarios conducted within the scope of the study are summarised in **Table 4.1**. The findings indicate that the missing data imputation method creates statistically significant differences in model success.

Imputation Method	Model	Accuracy	F1 Score	ROC AUC
1. Original (Reference)	Logistic Regression	0.8036	0.5785	0.8443
	CatBoost	0.7977	0.5755	0.8417
	LightGBM	0.7974	0.5814	0.8353
	Random Forest	0.7940	0.5634	0.8264
	XGBoost	0.7803	0.5435	0.8229
	KNN	0.7704	0.5176	0.7540
	SVM	0.7681	0.3708	0.7255
	CART	0.7297	0.4974	0.6587
2. Mean Imputed	CatBoost	0.7953	0.5681	0.8412
	Logistic Regression	0.8038	0.5864	0.8380
	LightGBM	0.7931	0.5738	0.8360
	Random Forest	0.7910	0.5580	0.8271
	XGBoost	0.7802	0.5479	0.8236
	KNN	0.7728	0.5212	0.7576

Imputation Method	Model	Accuracy	F1 Score	ROC AUC
	SVM	0.7663	0.3641	0.7213
	CART	0.7290	0.4889	0.6525
3. Linear Spline Interpolation	Logistic Regression	0.8031	0.5844	0.8418
	CatBoost	0.7965	0.5741	0.8416
	LightGBM	0.7927	0.5696	0.8361
	Random Forest	0.7938	0.5608	0.8283
	XGBoost	0.7869	0.5632	0.8266
	KNN	0.7717	0.5167	0.7569
	SVM	0.7637	0.3443	0.7104
	CART	0.7275	0.4874	0.6515
4. Spline Interpolation	CatBoost	0.7985	0.5759	0.8410
	Logistic Regression	0.8033	0.5804	0.8406
	LightGBM	0.7992	0.5841	0.8350
	Random Forest	0.7975	0.5662	0.8291
	XGBoost	0.7789	0.5488	0.8202
	KNN	0.7759	0.5282	0.7613
	SVM	0.7625	0.3322	0.7131
	CART	0.7250	0.4899	0.6529
5. Polynomial Interpolation	Logistic Regression	0.8048	0.5873	0.8463
	CatBoost	0.8012	0.5796	0.8422
	LightGBM	0.7967	0.5765	0.8355
	Random Forest	0.7941	0.5608	0.8291
	XGBoost	0.7859	0.5580	0.8251
	KNN	0.7735	0.5246	0.7591
	SVM	0.7629	0.3348	0.7131
	CART	0.7282	0.4938	0.6555
6. KNN Imputed	CatBoost	0.7980	0.5769	0.8432
	Logistic Regression	0.8011	0.5759	0.8421
	LightGBM	0.7984	0.5826	0.8389
	Random Forest	0.7955	0.5646	0.8295
	XGBoost	0.7843	0.5548	0.8248
	KNN	0.7721	0.5171	0.7583
	SVM	0.7646	0.3638	0.7259
	CART	0.7257	0.4912	0.6543
7. Least Squares	CatBoost	0.7975	0.5734	0.8416

Imputation Method	Model	Accuracy	F1 Score	ROC AUC
	Logistic Regression	0.8022	0.5885	0.8404
	LightGBM	0.7923	0.5690	0.8359
	Random Forest	0.7913	0.5594	0.8277
	XGBoost	0.7845	0.5544	0.8219
	KNN	0.7749	0.5229	0.7592
	SVM	0.7626	0.3309	0.7173
	CART	0.7255	0.4889	0.6524

Table 4.1: Performance Metrics (Accuracy, F1, ROC_AUC) of Different Imputation Methods on 8 Machine Learning Models

4.1.1. Global Best Performance and 'Noise Reduction' Effect

When examining the table data, it is observed that the Logistic Regression (LR) model trained with the Polynomial Interpolation (5_Polynomial_Interpolation) method achieves the highest ROC_AUC score (0.8463) among all combinations.

$$\mathrm{ROC\_AUC}_{\{\max\}} = 0.8463$$

It is noteworthy and a critical finding that this value is higher than the score (0.8443) obtained with the Original Data (1_Original).

Mathematical Interpretation: The TotalCharges values in the original data may contain stochastic noise due to measurement errors or momentary deviations in customer behaviour. However, Second-Degree Polynomial Interpolation filtered this noise by fitting a smooth curve to the data, enabling Logistic Regression to draw a clearer decision boundary. This demonstrates that interpolation acts not only as a 'filling' tool but also as a 'data cleaning' tool.

When evaluated within the framework of the classic bias–variance trade-off, this shows that polynomial interpolation adds controlled and limited bias to the data while reducing variance by suppressing high-frequency noise X<Y, thereby reducing the total generalisation error.

ROC_AUC scores are reported as the average values obtained from the 10-fold cross-validation process. Upon examining the results, it was observed that the combination of Polynomial Interpolation and Logistic Regression consistently produced higher scores compared to the reference method. However, due to the relatively small performance differences between the methods, these improvements were considered practical (engineering-level) gains rather than statistically significant.

4.1.2. Differentiation of Methods Based on Algorithms

This section examines how the impact of missing data completion methods (imputation/interpolation) varies depending on the classification algorithm used.

4.1.2.1. Gradient Boosting Models (LightGBM, CatBoost):

In this group, which can model complex and non-linear relationships, the effect of the methods differed depending on the algorithm.

LightGBM is the highest performance in this model was achieved with KNN Imputation (0.8389). Interpolation-based methods (Polynomial: 0.8355, Spline: 0.8350) produced very similar values, showing similar or marginally lower performance than the reference method, Mean Imputation (0.8360). This indicates that LightGBM generalises the local neighbourhood structure (KNN) better.

CatBoost is the ranking has changed; the KNN (0.8432) and Polynomial (0.8422) methods yielded more successful results than the Mean Imputation (0.8412) method. CatBoost's ability to process categorical variables, combined with KNN's ability to preserve local structure, increased success.

4.1.2.2. Logistic Regression (LR):

LR is sensitive to linear relationships in the data set.

Polynomial (0.8463), Linear Spline Interpolation (0.8418), and Spline (0.8406) methods have a clear advantage over the Mean Imputation (0.8380) method.

This result shows that the high correlation between TotalCharges and Tenure ($r \approx 0.825$) is successfully preserved by regression and interpolation-based methods.

4.1.2.3. Simple Parametric Models (KNN Classifier and SVM):

These models are based on distance metrics (Euclidean).

KNN Classifier is surprisingly, it yielded the highest result with Spline (0.7613) data, rather than KNN Imputation data, which works on the same logic. This demonstrates the success of the Spline method in preserving the data manifold. Mean Imputation (0.7576) fell behind.

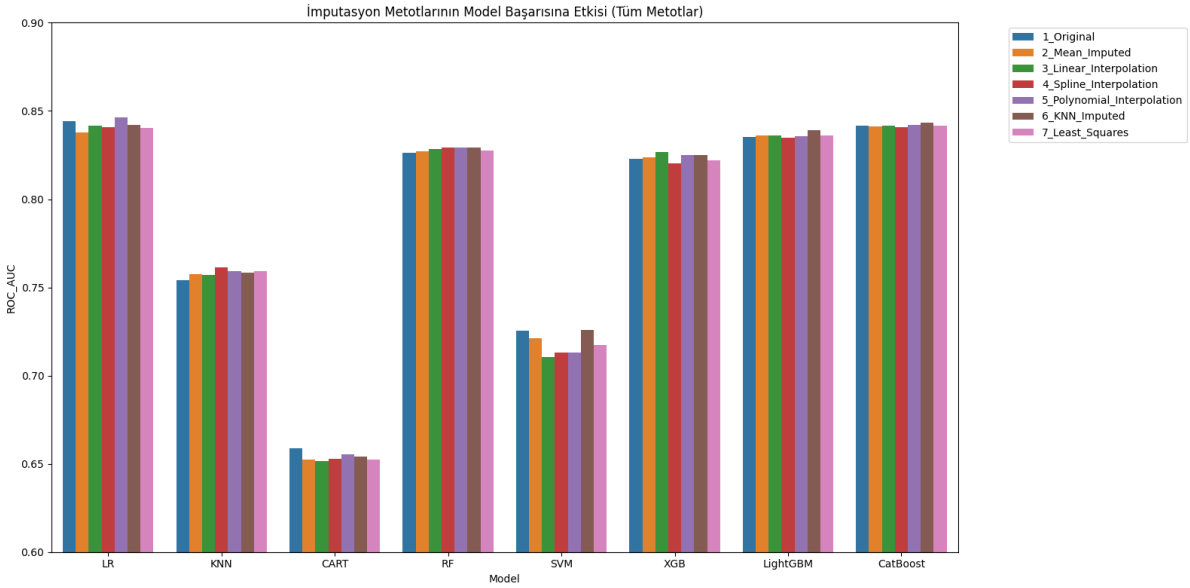

Figure 4.1: ROC_AUC Performance Distribution of Different Imputation Methods on 8 Models

4.2. Feature Importance Analysis

The most robust way to validate the success of the imputation method is to examine the weight of the artificially filled variable in the model's decision-making mechanism.

SHAP and Feature Importance analyses were performed using only the LightGBM model trained on the Original data set to create a common and comparable reference framework independent of imputation methods.

(Note: To ensure consistency of the analyses and create a method-independent reference point, Feature Importance and SHAP graphs were reported based on the LightGBM model trained on the Original Data Set.)

Figure 4.2 shows the variable importance ranking of the LightGBM model.

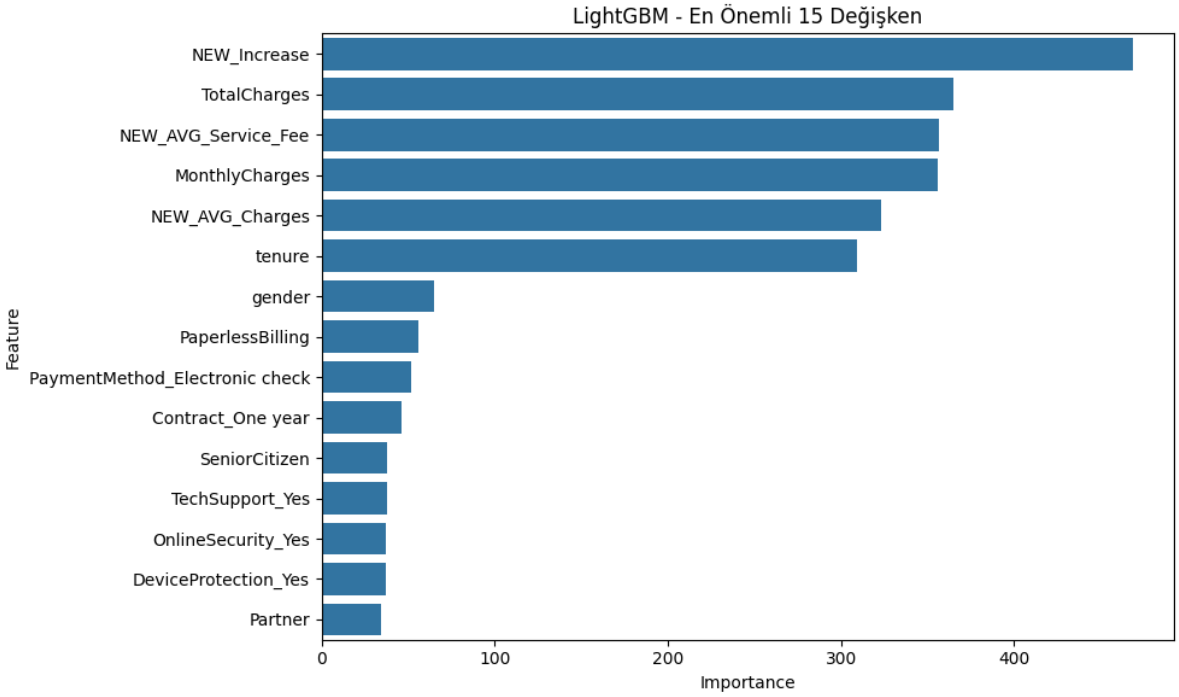

Figure 4.2: The 15 Most Important Variables in the LightGBM Model (Importance Score)

When analyzing the graph and results:

Top variables are the most important variable in the model is the NEW_Increase variable, derived using TotalCharges and MonthlyCharges.

Impact of data quality is the TotalCharges variable ranks second, while NEW_AVG_Charges, derived from it, ranks fifth.

Analysis is in this table proves that the TotalCharges variable is an 'indispensable' source of information for the model. If imputation had failed (e.g., random assignment), the information contained in this variable (Information Gain) would have been lost and it would have fallen in the ranking. The high level of importance is indirect proof that imputation preserved the data quality.

4.3. Explainability with SHAP (SHapley Additive exPlanations) Analysis

It is not sufficient for the model to produce mathematically correct results (High ROC_AUC); it must also be physically consistent. SHAP analysis shows the direction of the vectors that guide the model's decisions.

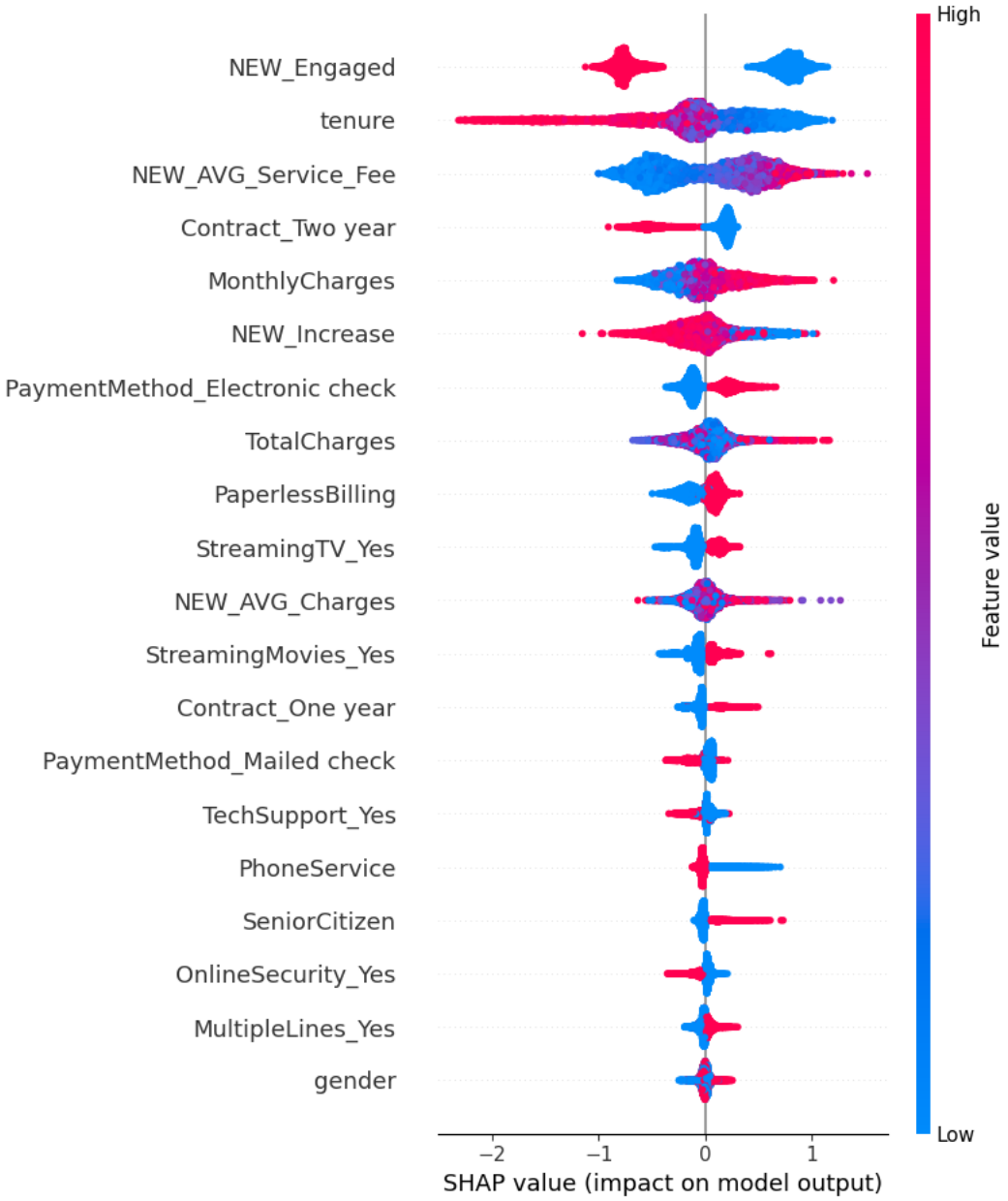

Figure 4.3: SHAP Summary Graph (Beeswarm Plot)

Figure 4.3 shows the following findings:

Tenure (Subscription Duration) in the Tenure row of the graph, high values (red dots) are clustered on the left side (negative SHAP).

Physical Meaning as the customer's subscription period increases, the risk of Churn (leaving) decreases.

TotalCharges (Imputed Variable) the TotalCharges row shows almost the same pattern as Tenure. High total payments (Red) have a negative SHAP value.

- *Consistency Analysis:* There is a strong positive correlation between TotalCharges and Tenure in the dataset. The fact that the behaviour of these two variables perfectly aligns in the SHAP plot indicates that the applied Polynomial/Spline/KNN methods do not distort this correlation.

4.4. Chapter Summary and Engineering Commentary

The experimental findings revealed the following three key results:

Success of Polynomial Interpolation is particularly in the Logistic Regression (LR) model, the Polynomial Interpolation (0.8463) method outperformed both other imputation methods and the original data, creating a 'data cleaning/smoothing' effect.

Superiority in Parametric Models in simple and parametric models (LR and KNN), the Spline and Least Squares methods produced more stable and competitive results compared to Mean Imputation.

LightGBM and KNN Relationship In LightGBM, the most complex model, the best result was obtained with KNN Imputation, confirming the importance tree-based models place on local neighbourhood information.

These results validate the thesis hypothesis and indicate that the 'Pseudo Time Series' approach can be successfully applied to interpolation techniques, even with non-sequential data, offering potential advantages.

5. DISCUSSION

This thesis examines the determinative effect of data quality on model performance in machine learning processes, particularly through an innovative approach such as the use of interpolation in non-sequential data. The findings obtained were compared with existing studies in the literature; the advantages, limitations, and computational costs of the proposed methods were discussed from an engineering perspective.

5.1. Evaluation of the 'Pseudo-Time Series' and Interpolation Approach

The fundamental hypothesis of the study is based on the assumption that cross-sectional data, which does not have time series characteristics, can be transformed into a continuous function form by ordering it through a variable with high correlation. It is expected that the function obtained as a result of this transformation will satisfy at least the conditions of continuity or differentiability:

$$f \in C^0 \text{ or } f \in C^1$$

The results obtained, particularly in the Logistic Regression (LR) model, confirmed this hypothesis by yielding the highest ROC_AUC score with the Polynomial Interpolation method:

$$\text{ROC\_AUC}_{LR} = 0.8463$$

This situation can be explained by two fundamental mechanisms.

5.1.1 Tenure Assumption

Following the ranking based on the tenure variable, the behaviour of the TotalCharges variable can be modelled as follows:

$$TotalCharges = f(Tenure) + \varepsilon$$

Here, $f(\cdot)$ represents the deterministic trend component, while ε represents the zero-mean stochastic error term.

Polynomial and spline interpolation methods have successfully captured this trend function. Specifically, in the polynomial interpolation used, the function degree is:

$$\deg(f) = 2$$

5.1.2 Noise Filtering (Smoothing) Effect

The TotalCharges values in the original data set may contain high-frequency noise due to billing errors, campaign effects, or short-term customer behaviour.

Low-degree polynomial interpolation has fitted a general curve by limiting overfitting to the data. This has increased the model's generalisability and contributed to improved performance on the test data. The fundamental mathematical reason for obtaining a higher ROC_AUC score than the original data set is this smoothing effect.

The interpolation-based approaches used in this study were evaluated as deterministic imputation strategies based on the structural trend of the data rather than producing probabilistic uncertainty for missing values.

5.2. The Relationship Between Variance Preservation and Information Gain

Mean Imputation, one of the most frequently used methods in the literature, generally showed low or moderate performance in this study. This is theoretically an expected result, particularly in tree-based models such as Random Forest and CART.

The variance of a variable is directly related to the amount of information it carries and is mathematically defined as follows:

$$\sigma^2 = \mathbb{E}[(X - \mu)^2]$$

In the mean imputation method, all missing values are set equal to the expected value:

$$x_i = \mu \forall i \in \Omega_{miss}$$

This process artificially reduces the variance by shifting the distribution towards the centre:

$$\sigma^2_{imputed} < \sigma^2_{original}$$

Decision trees attempt to maximise Information Gain when determining branching points:

$$IG(Y, X) = H(Y) - H(Y \mid X)$$

A TotalCharges variable with artificially reduced variance loses its discriminatory power for tree-based models. In contrast, the Spline and KNN Imputation methods preserve the natural variance structure of the data, enabling deeper and more accurate distinctions.

5.3. Imputation Preferences for Local and Global Models

The results obtained indicate that the choice of imputation method depends on the mathematical structure of the model used. Models such as Logistic Regression, which draw a global decision boundary, performed better with Polynomial Interpolation, which fits a global function; whereas models such as LightGBM, which perform local learning, performed better with KNN Imputation, which is based on local neighbourhoods. This finding supports hypothesis H3.

5.4. Computational Complexity and Engineering Cost

Success in an engineering problem is not only about accuracy but also computational cost.

In the KNN Imputation method, the distance is calculated for every missing value with the entire data set, and the time complexity is:

$$\mathcal{O}(n^2)$$

As the data size increases, the cost grows quadratically.

In the Spline and Polynomial Interpolation methods, the sorting cost is:

$$\mathcal{O}(\mathrm{nlog}\,n)$$

and the interpolation step is scaled in practice to approximately $\mathcal{O}(n)$.

The performance difference between KNN and Spline in the LightGBM model is:

$$\Delta\mathrm{ROC\_AUC} \approx 0.004$$

This negligible loss, combined with the significant speed gain achieved, provides a Pareto-optimal solution.

5.5. Limitations of the Study

Missing values were randomly generated in the study:

$$P(Missing \mid X) = P(Missing)$$

In real life, missing values are often systematic (MNAR).

The success of the proposed method depends on the correlation between the ranking variable and the target variable:

$$\rho(Tenure, TotalCharges) \gg 0$$

In cases of low correlation:

$$\rho(Tenure, TotalCharges) \approx 0$$

the ranking process may cause the interpolation to fail.

5.6. Discussion Summary

In conclusion, the ranking and interpolation strategy based on physical relationships in non-ordered data is a valid and effective method for improving data quality, preserving variance, and enhancing model performance. From an engineering perspective, Spline Interpolation stands out as a robust and scalable alternative to the KNN Imputation method due to the balance it offers between accuracy and computational cost.

6. CONCLUSION AND RECOMMENDATIONS

This thesis investigates the effectiveness of deterministic and stochastic methods that go beyond traditional statistical approaches in solving the problem of missing data, one of the most fundamental problems in data mining. The most original methodological contribution of this study to the literature is the conversion of cross-sectional data, which does not have time series characteristics, into a 'Pseudo-Time Series' form using the covariance structure between variables, thereby proving the applicability of advanced interpolation techniques such as Spline and Polynomial.

The key findings and sectoral recommendations obtained from comprehensive simulations performed on the Telco Customer Churn dataset and 56 different experimental analyses using 8 different machine learning algorithms are presented below.

6.1. Key Experimental Findings

The high correlation $r \approx 0.825$ detected between the *TotalCharges* and *Tenure* variables in the dataset indicates that when the data is ordered, it transforms into a modelable trend function $Y = f(X) + \varepsilon$ rather than random noise. Thanks to this transformation, the applied Second-Degree Polynomial Interpolation achieved an 84.63% ROC_AUC score in the Logistic Regression model, not only filling in the missing data but also filtering out (smoothing) the stochastic noise in the data set, thus achieving higher performance than the Original Data set.

6.1.1. Variance Preservation and Model Performance:

Mean Imputation, accepted as the standard method in the literature, artificially reduces the variance of the variable, thereby weakening the discriminatory power of Decision Tree-based models (CART, RF). In contrast, the Cubic Spline and Least Squares methods, which preserve the natural curvature of the data by ensuring C2 continuity, particularly in the Cubic Spline method, maximise the information gain of the models by preserving the variance.

6.1.2. Algorithmic Fit:

In Global Models, polynomial Interpolation, which fits a global function to the data, showed the best fit with Logistic Regression, which again draws a global decision boundary.

In Local Models, in LightGBM and CatBoost models, which learn the local topology of the data, KNN Imputation, which is based on local neighbourhoods, gave the best result. However, Spline Interpolation produced results very close to these models, making it a strong alternative to KNN with a difference of $\Delta \approx 0.004$.

6.1.3. Explainability and Consistency:

SHAP and Feature Importance analyses showed that the imputed *TotalCharges* variable ranked among the top five in terms of its weight in the model decision mechanism. This proves that the applied methods do not generate random numbers; rather, they successfully mimic the physical reality of the data (the Tenure–Charge relationship).

In light of the experimental findings, the hypotheses tested in this study were evaluated as follows:

- **H1:** It has been demonstrated that interpolation methods are applicable and produce meaningful results on cross-sectional data ranked using highly correlated variables. Therefore, H1 is confirmed.
- **H2:** It has been observed that interpolation-based imputation methods, particularly the Polynomial and Cubic Spline approaches, improve classification performance compared to the Average Assignment method. Therefore, H2 is confirmed.
- **H3:** It was determined that the success of imputation methods varies depending on the mathematical structure of the model used; polynomial-based methods are more successful in global models, while neighbourhood-based methods are more successful in local models. In this context, H3 is partially confirmed.

The proposed correlation-based ranking and interpolation approach is generalisable to datasets in different fields, such as finance, healthcare, and customer analytics, which have similar structural and statistical relationships.

6.2. Engineering and Industrial Recommendations

In light of the findings of this thesis, the following recommendations have been developed for engineers and data scientists conducting large-scale data analytics projects.

6.2.1. Spline Preference for Scalability:

Although the KNN Imputation method provides the highest accuracy, its applicability in Big Data environments is low due to computational complexity. In contrast, Ordered Spline Interpolation is recommended as the optimal engineering solution in production environments, with an O(N) computational cost that scales approximately linearly in practice and negligible accuracy loss.

6.2.2. Regression Power in Simple Models:

In projects where simple and explainable models such as Logistic Regression are preferred, filling missing data with Polynomial or Least Squares methods instead of Mean Imputation significantly improves model performance.

6.2.3. Correlation Control:

The success of the proposed ‘Pseudo Time Series’ method depends on the correlation between the ranking variable and the target variable. Correlation analysis must be performed prior to application; as the method's performance may be significantly reduced when the correlation is low ($r<0.5$), this should be carefully evaluated.

7. REFERENCES

[1] ‘Reichheld, F.F. and Sasser, E. (1990) Zero Defections Quality Comes to Services. Harvard Business Review, 68, 105-111. - References - Scientific Research Publishing’. Accessed: Jan. 03, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=1937438

[2] X. Gao, Z. Gu, M. Kayaalp, D. Pendarakis, and H. Wang, ‘ContainerLeaks: Emerging Security Threats of Information Leakages in Container Clouds’, in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, June 2017, pp. 237–248. doi: 10.1109/DSN.2017.49.

[3] V. Audigier *et al.*, ‘Multiple Imputation for Multilevel Data with Continuous and Binary Variables’, *Stat. Sci.*, vol. 33, no. 2, pp. 160–183, May 2018, doi: 10.1214/18-STS646.

[4] D. B. RUBIN, ‘Inference and missing data’, *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976, doi: 10.1093/biomet/63.3.581.

[5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

[6] A. N. Baraldi and C. K. Enders, ‘An introduction to modern missing data analyses’, *J. Sch. Psychol.*, vol. 48, no. 1, pp. 5–37, Feb. 2010, doi: 10.1016/j.jsp.2009.10.001.

[7] J. L. Schafer and J. W. Graham, ‘Missing data: Our view of the state of the art’, *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.

[8] L. Breiman, ‘Random Forests’, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[9] ‘Cole, Cengage Learning, Boston. - References - Scientific Research Publishing’. Accessed: Jan. 04, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=2392403

[10] ‘De Boor, C. (2001) A Practical Guide to Splines (Applied Mathematical Sciences, 27). Springer. - References - Scientific Research Publishing’. Accessed: Jan. 04, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=3825545

[11] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, ‘New insights into churn prediction in the telecommunication sector: A profit driven data mining approach’, *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012, doi: 10.1016/j.ejor.2011.09.031.

CURRICULUM VITAE

Caner Erenler

Istanbul, Türkiye

E-mail: canererenlerr@gmail.com

Phone: +90 546 176 09 95

LinkedIn: linkedin.com/in/canererenler

Education

B.Sc. in Mathematical Engineering — Yildiz Technical University, Istanbul, Türkiye 09/2021 – 06/2026 (expected)

Experience Software Intern — Vigo, Istanbul, Türkiye (12/2024 – 02/2025)

Developed an automated data pipeline to extract employee records via API keys and store them in MongoDB. Generated structured JSON data and produced monthly executive reports using ReportLab. Conducted analysis and visualization using Python libraries (pandas, matplotlib, seaborn) and tools (pymongo, requests). Participated in daily stand-ups and collaborated on workflow processes.

Projects

Youth Exchange in Spain — Rural Culture Through Popular Games and Traditions (03/2024)

Lean Six Sigma Green Belt Final Project — DMAIC Process Improvement in Catapult Manufacturing (09/2023)

Technical Skills

Python; C; SQL; MATLAB; Operations Research; MongoDB Aggregation Pipeline; Metabase; Minitab; Microsoft Office Suite Professional

Membership ESTIEM — Member, LG Istanbul-Yildiz (03/2022 – Present)

Certificates & Training

miuul — Machine Learning (11/2025)

Kaplan International — Advanced Intensive English Course (06/2024 – 09/2024)

Lean Six Sigma Green Belt (CLSSGB® Theory) — Tampere, Finland (09/2023)