# D208 Predictive Modeling linear regression

Can service-related factors like monthly charges, bandwidth usage, and customer interactions predict customer tenure? **Tenure** is our dependent variable, and our independent variables for this analysis are **MonthlyCharge**, **Bandwidth\_GB\_Year,** **Outage\_sec\_perweek**, **Contacts**, **Yearly\_equip\_failure, Contract**, **InternetService**, **PaymentMethod**, **TechSupport**, and **OnlineSecurity**. The goal of the data analysis question is to understand the factors that most influence customer tenure. This allows the business to focus on the most impactful areas for intervention. The purpose of this model is to be proactive in identifying areas of improvement, so we can anticipate customer needs and allocate resources to lengthen their time with us.

**Method Justification**

Linear regression is suitable for this analysis as it allows use to quantify the impact of each service-related factor on our continuous dependent variable(tenure), while considering potential linear relationships. Linear regression is better suited for this analysis because our dependent variable is continuous rather than categorical. Logistic regression would be better suited if we were looking to predict a categorical dependent variable. When using a multiple linear regression model, several key assumptions must be met to ensure that the model is valid, and the results are reliable. The relationship between the independent variables and the dependent variable is linear. The data variables are independent of each other. The variance of the residuals is constant across all levels of the independent variables. The differences between the observed and predicted values are normally distributed.

The comprehensive data handling and preprocessing Python provides is why I prefer it for data analysis. It offers powerful libraries like **Pandas** and **NumPy** for data manipulation, cleaning, and transformation, which are essential in the preprocessing phase of analysis. Python supports a wide range of statistical and machine learning libraries such as **statsmodels**, **scikit-learn**, and **SciPy**, which are invaluable for developing, validating, and interpreting multiple linear regression models.

**Data Preparation**

My data cleaning goals are to ensure data quality, address outliers and anomalies, and prepare data for modeling. I first load in and inspect the data using **df = pd.read\_csv()**. I want to understand the structure of the dataset, including the types of variables and the presence of any obvious data quality issues, so I use **df.info()**, **df.describe()**, **df.head()**, and **df.dtypes** to get more insight into the dataset. I then go into identifying missing values using **df.isnull().sum()**. I decide to impute them or remove the rows/columns in most cases, but my dataset doesn’t have any missing values. I also check for duplicates in this phase as well using **df.duplicated().sum()**.

Formatting and standardizing my data is my next focus. I ensure that all variables are in the correct format for analysis, converting data types if necessary. My selected variables are numerical and categorical, I’ve encoded my categorical variables and I’ve also addressed outliers that could disproportionately influence the results of the regression model for my continuous. Outliers were replaced with median values to minimize their impact on the regression model while preserving the central tendency of the data. Once all this is done, I Finalize the dataset, ensuring all necessary transformations are complete, and the data is in the correct format for regression analysis.

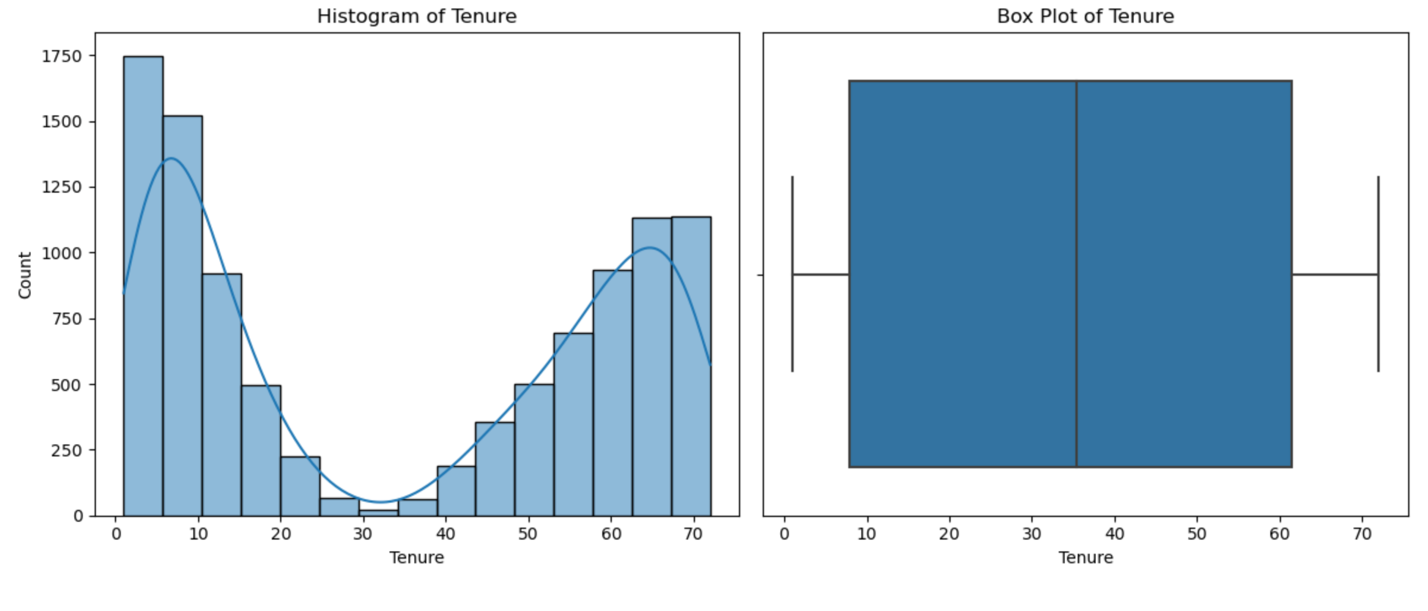
**Tenure** our dependent variable represents the number of months a customer has been with our company. We want to know if we can predict a customer’s tenure based on their service data.

A number of numbers and symbols

Description automatically generated with medium confidence

A screenshot of a calculator

Description automatically generated



A graph of a diagram

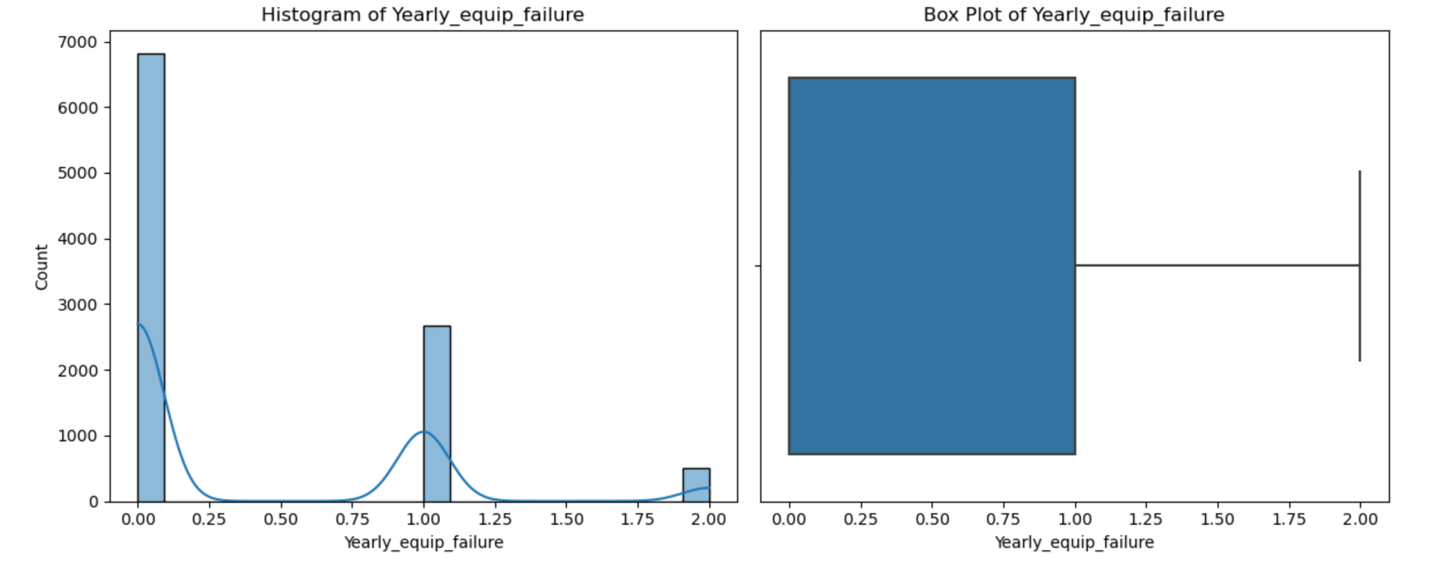
Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

A graph of a function

Description automatically generated

A graph of contact information

Description automatically generated with medium confidence

A graph of blue rectangular bars

Description automatically generated with medium confidence

A number and numbers on a white background

Description automatically generated

A graph of blue rectangular bars

Description automatically generated with medium confidence

A close-up of numbers

Description automatically generated

A graph of a distribution of payment method

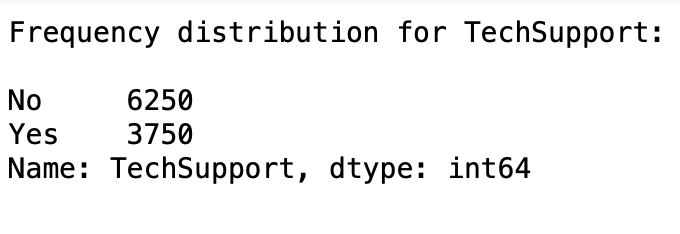
Description automatically generated

A white background with black text

Description automatically generated

A graph with blue rectangles

Description automatically generated



A graph with blue rectangular bars

Description automatically generated with medium confidence

A black text on a white background

Description automatically generated

A diagram of a graph

Description automatically generated with medium confidence

A number and numbers on a white background

Description automatically generated

A diagram of different colored squares

Description automatically generated

A screenshot of a computer

Description automatically generated

A diagram of different colored rectangular shapes

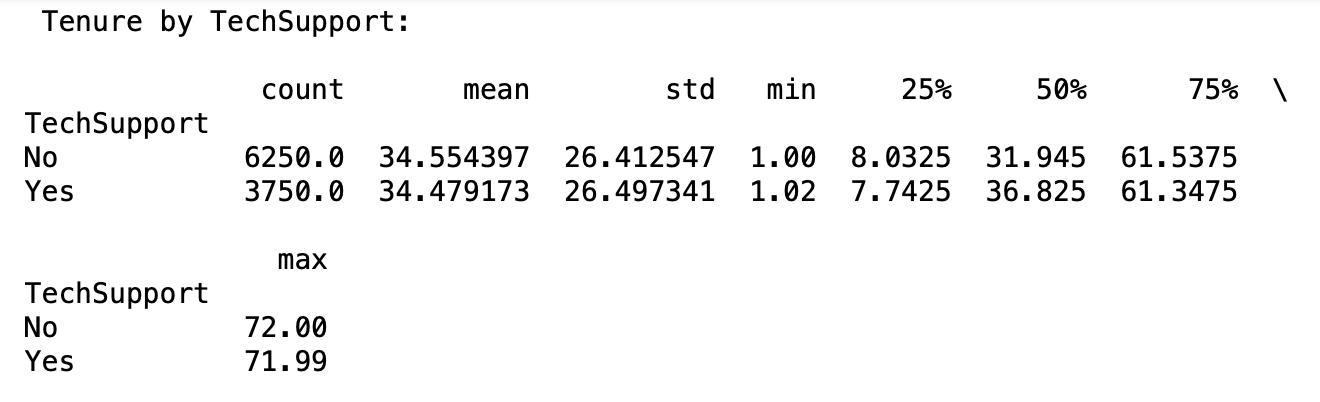
Description automatically generated

A screenshot of a white sheet with numbers

Description automatically generated

A diagram of a diagram

Description automatically generated with medium confidence



A diagram of a chart

Description automatically generated with medium confidence

A number and numbers on a white background

Description automatically generated

A graph of a line

Description automatically generated

A graph with a red line

Description automatically generated

A graph with a red line

Description automatically generatedA diagram of a graph

Description automatically generated with medium confidenceA graph showing the difference between tenue and bandwidth

Description automatically generatedA diagram of a bar code

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

In summary I’ve loaded in the dataset and cleaned the data by making sure there are no duplicates or missing values. I’ve accounted for outliers by replacing them with the columns median values and I’ve gathered statistical summaries for the chosen variables. Visualizations for univariate and bivariate distributions have also been provided. Moving into modeling I have a dataset of my chosen features that are evenly distributed and encoded to align with regression modeling so that we may get an accurate model for predicting customer tenure.

**Model Comparison & Analysis**

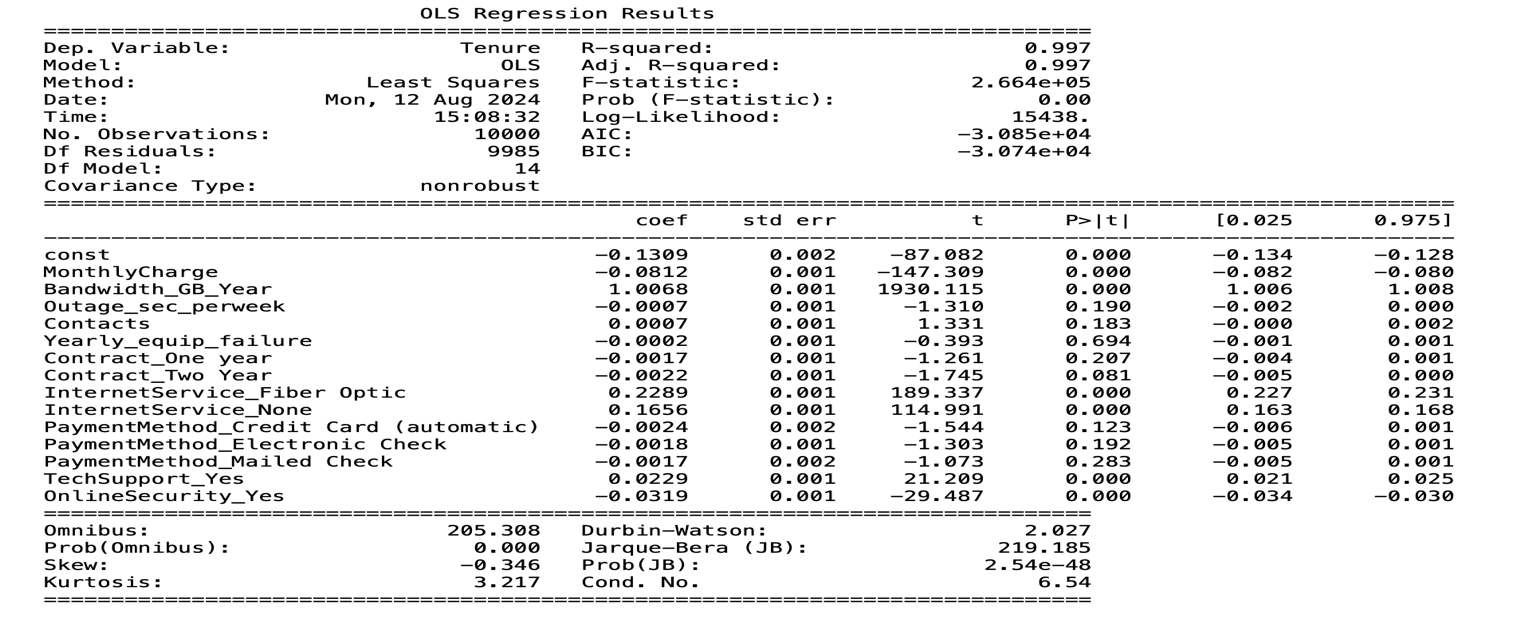
I used the **OLS** function from the **statsmodels** library to fit the multiple linear regression model. I then provided the summary of the model to provide insight on how each of the variables influenced ‘**Tenure**’. I used **Backward Elimination** to verify the statistical significance of my chosen features. Backward elimination was selected for its efficiency in removing non-significant variables, which helped streamline the model to include only those predictors with the strongest statistical significance. This process uses p-value above the .05 threshold to help you remove less significant variables. In this case our model features dropped from fourteen down to six significant features. Adjusted R-squared is used for my model evaluation. It allows for the comparison of models with different numbers of predictors, helping to identify the model that best balances complexity and performance. The model with the highest adjusted R-squared, indicates the best fit with the least complexity.

**Model Statistics**

**A screenshot of a computer screen

Description automatically generated**

**Regression Model**

****

**Regression Model/After Backwards Elimination**

**A screenshot of a computer

Description automatically generated**

**Residual Plot**

**A graph showing a number of blue dots

Description automatically generated**

**A graph of a graph

Description automatically generated**

**Data Summary & Implications**

My initial model includedfourteen features, but my reduced model used less variables based on the significance of each variables p-value. The reduced model has a p-value threshold of 0.05 so any feature with a higher p-value than the threshold is removed from the model. The equation below represents how each feature in the reduced model contributes to predicting the tenure of a customer. Tenure( -1.6940) is the baseline value of tenure when all other independent variables are set to zero. However, since some variables like InternetService\_Fiber Optic and TechSupport\_Yes are binary (1 or 0), the intercept by itself may not be directly meaningful in isolation. It essentially serves as a starting point for the model. MonthyCharge(-0.0500) means for every additional unit increase in monthly charge, tenure decreases by 0.05 months. This suggests that higher monthly charges are associated with shorter customer tenure. Customers tend to stay longer with lower charges.

Bandwidth\_GB\_Year(0.0122) means for every additional gigabyte of bandwidth used per year; tenure increases by 0.0122 months. This suggests that customers with higher bandwidth usage tend to stay longer, possibly because they are more engaged with the service. InternetService\_Fiber Optic (6.0514) means customers who use fiber optic internet tend to stay approximately six months longer compared to those who don't. This indicates that fiber optic service is a strong factor in retaining customers. InternetService\_None (4.3759) means if a customer does not have an internet service, their tenure increases by about 4.38 months.

This might indicate that customers using other services (such as TV or phone) without internet tend to stay longer. TechSupport\_Yes (0.6054) means customers who have tech support tend to stay an additional 0.61 months compared to those who don’t. This suggests that tech support availability positively impacts customer retention. OnlineSecurity\_Yes (-0.8433) means customers who have online security services tend to have a tenure that is approximately 0.84 months shorter than those without it. This could imply that online security, while a valued service, might not be a strong factor in customer retention, or customers might be switching to competitors offering better alternatives. MonthlyCharge and OnlineSecurity\_Yes have negative coefficients, meaning they reduce tenure, while Bandwidth\_GB\_Year, InternetService\_Fiber Optic, InternetService\_None, and TechSupport\_Yes have positive coefficients, meaning they increase tenure.

Tenure = -1.6940 – 0.0500 x MonthlyCharge + 0.0122 x Bandwidth\_GB\_Year + 6.0514 x InternetService\_Fiber\_Optic+4.3759 x InternetService\_None + 0.6054 x TechSupport\_Yes -0.8433 x OnlineSecurity\_Yes

The reduced regression model shows a strong statistical significance, with an R-squared value of 0.997, indicating that 99.7% of the variance in customer tenure is explained by the model’s predictors. The F-statistic of 6.214e+05(621,400)and its associated p-value of 0.00 confirm that the model is highly significant. Additionally, all independent variables retained in the model—such as MonthlyCharge, Bandwidth\_GB\_Year, InternetService\_Fiber Optic, InternetService\_None, TechSupport\_Yes, and OnlineSecurity\_Yes have p-values of 0.00, meaning they are each individually significant in predicting customer tenure. Overall, the model demonstrates a strong fit and meaningful statistical significance in explaining customer tenure.

Practical significance suggests even if a coefficient is statistically significant, the magnitude of the coefficient matters for practical significance, even though some variables like MonthlyCharge have a statistically significant impact on tenure, the magnitude of their effect is small (0.05 months). This may be statistically significant but not practically impactful in the sense that small pricing changes might not noticeably affect customer behavior. Variables like InternetService\_Fiber Optic and TechSupport\_Yes are practically significant because they contribute several months of tenure, which can have a substantial effect on overall customer retention strategies.

**Limitations of Data Analysis**

The model assumes linear relationships between the independent variables and tenure. If the true relationships are nonlinear or more complex, this model might not fully capture those dynamics. The condition number in the results suggests potential multicollinearity. This means some of the independent variables may be correlated, which can affect the reliability of coefficient estimates. Variables like bandwidth, internet service type, and tech support might have underlying correlations that affect their individual contributions. While the model explains a large portion of the variance, there could be important factors not captured by the data, leading to omitted variable bias. Factors like customer demographics or satisfaction surveys could potentially improve the model. The model assumes normally distributed errors and consistent variance across the dataset (homoscedasticity). If these assumptions are violated, the model's predictive power could be compromised.

**Recommended Actions**

Given that **MonthlyCharge** has a negative impact on tenure, the company should reconsider its pricing strategy by offering discounts or incentives to high-bandwidth customers or those at risk of leaving, this could help extend their tenure. The model shows that **InternetService\_Fiber Optic** has a strong positive impact on tenure. Expanding the availability of fiber optic internet could be a strategic priority to retain more customers. **TechSupport\_Yes** is associated with increased tenure, offering tech support as an add-on service or improving its quality could help retain customers. Additionally, marketing tech support as a key feature may attract longer-term customers. Interestingly, **OnlineSecurity\_Yes** is associated with a decrease in tenure. This suggests that customers might not perceive enough value in this service, or they may be opting for better alternatives. The company should investigate customer perceptions of this service and consider either improving it or offering better alternatives.

The company can identify customers predicted to have shorter tenure (based on the model) and target them with retention offers such as loyalty rewards, discounted services, or bundled packages that include tech support or fiber optic internet. The model should be updated regularly with new data to maintain accuracy and reflect changes in customer preferences or external factors (e.g., competitors' services). Additionally, collecting more customer behavior data, such as satisfaction scores or engagement metrics, could further enhance predictive accuracy. By focusing on service areas that significantly increase customer tenure and addressing the negative impacts of high charges and underperforming services, the company can better retain customers and improve long-term profitability.

**References**

groverjatin.medium.com, accessed 06 August 2024, https://groverjatin.medium.com/short-python-code-for-backward-elimination-with-detailed-explanation-52894a9a7880

scikit-learn.org, accessed 06 August 2024, https://scikit learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

www.statology.org, accessed 06 August 2024, https://www.statology.org/standardized-residuals-python/

sphweb.bumc.bu.edu, accessed 06 August 2024, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713\_multivariablemethods/bs704-ep713\_multivariablemethods2.html

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy-1.6.3/reference/generated/scipy.stats.ttest\_ind.html.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html#skew.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html#kurtosis.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html#scipy.stats.chisquare.

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA) accessed 09 June 2024, https://apps.cgp-oex.wgu.edu.

Scikit-learn, accessed 09 June 2024, <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/numpy/numpy\_array\_iterating.asp.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/pandas/default.asp.