Results Report

INTERCONNECT

Martha Patricia Ortiz

Final Project

tripleten

February 2025

INTERCONNECT

Background

If a user is identified as planning to leave, they will be offered promotional codes and special plan options. Interconnect's marketing team has gathered some customer data, including information about their plans and contracts.

Goal

Interconnect aims to forecast its customer churn rate.

Work Plan

1. Initialization
2. Data Preprocessing
3. Exploring Data Analysis
   - Graphs
   - Levene Test
   - ANOVA Test
   - EDA Conclusions
4. Data Models
   - Dummy Model
   - HOE Models
     - Logistic Regression
     - Tree Classifier
     - Random Forest
   - Ordinal Encoding Model
     - Tree Classifier
     - Random Forest
     - Light GBM Classifier
     - Cat Boost Classifier

1. Inicialization

Libraries:

- math: Provides basic mathematical functions.
- numpy (np): Library for numerical operations and handling multidimensional arrays.
- pandas (pd): Tool for data manipulation and analysis using DataFrames.
- seaborn (sns) and matplotlib.pyplot (plt): Used for data visualization and generating plots.
- scipy.stats (st): Offers statistical tools and functions for data analysis.

Modeling and Machine Learning

- lightgbm (lgb): Optimized library for Gradient Boosting models.
- catboost (CatBoostClassifier): Boosting-based model optimized for categorical data.

- sklearn (Scikit-learn):
  - Models:
    - DummyClassifier
    - DecisionTreeClassifier
    - RandomForestClassifier
    - LogisticRegression
    - LinearRegression.

- Metrics:
  - accuracy_score
  - f1_score
  - log_loss
  - roc_auc_score
- Preprocessing:
  - OrdinalEncoder
  - StandardScaler
  - shuffle

- Data Splitting: train_test_split

This set of tools enables exploratory data analysis, preprocessing, classification and regression modeling, as well as model evaluation.

2. Data Preprocessing

Data Sets:

- df_contract: Information about the contract signed by the customer.
- df_personal: Personal information of the customer.
- df_internet: Information about the internet services contracted.
- df_phone: Indicates if the customer subscribed to multiple lines.

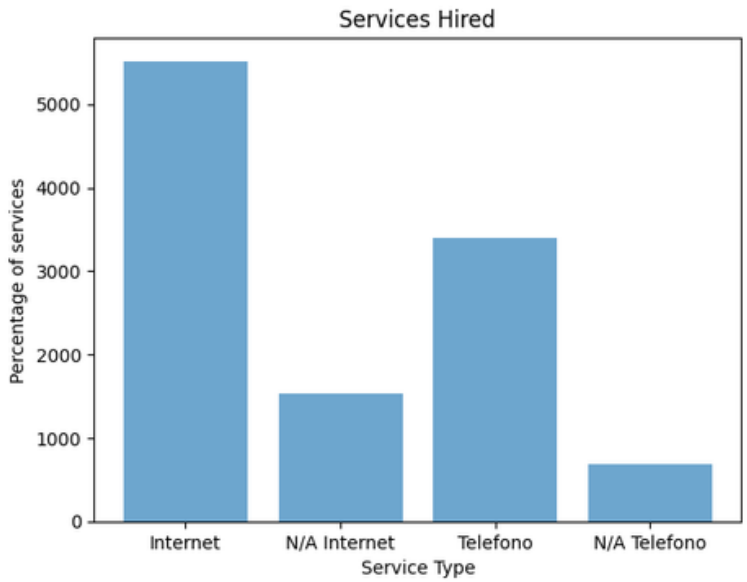
Modeling and Machine Learning

The four datasets were reviewed to ensure the information was complete, with no duplicate or missing data. Column data types were reassigned to integers where necessary.

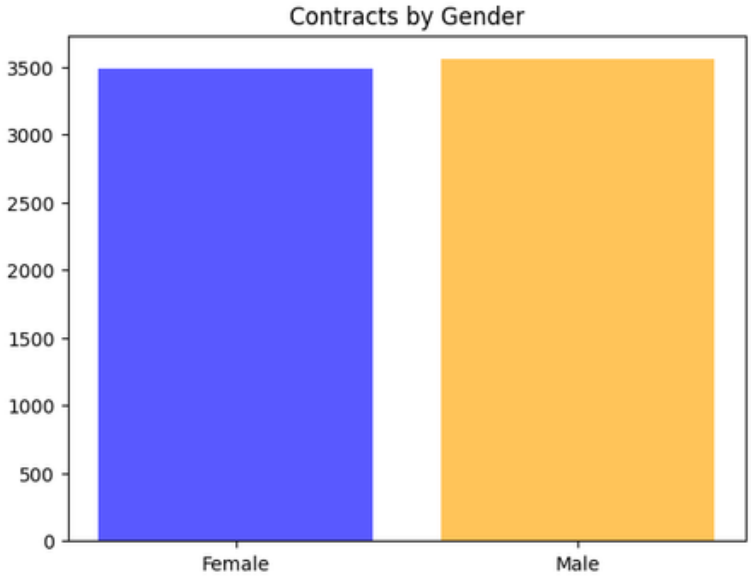
Columns names were converted to lowercase.

The four datasets were merged into a single table to consolidate the information.

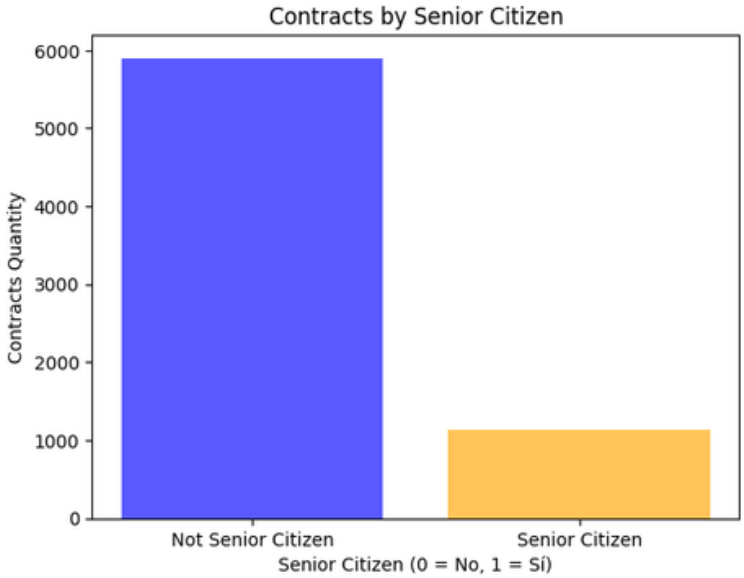
3. Exploring Data Analysis



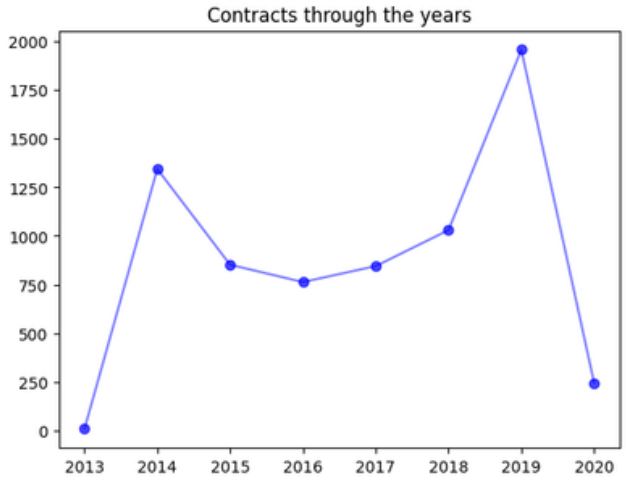
Most users subscribe to internet services more than phone services.

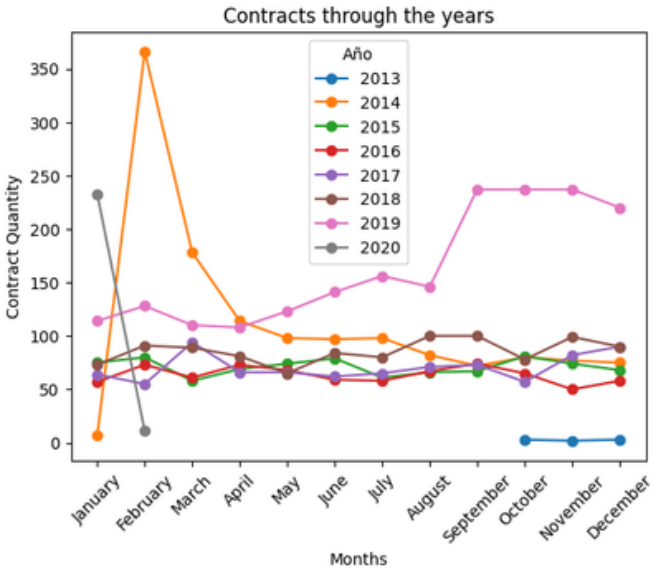


The number of male and female users is very similar.



The majority of users are under 60 years old, suggesting that they are young individuals.





Yearly Contract Trends:

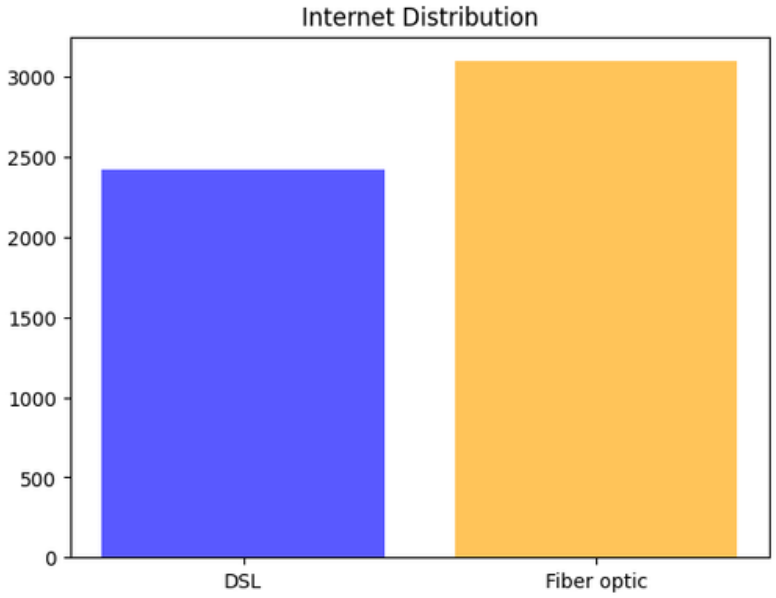
- The number of contracts fluctuates over the years, with a significant peak in 2019, showing the highest contract volume.
- There is a notable drop in 2020, which could be due to external factors such as economic downturns or global events.

Monthly Contract Trends:

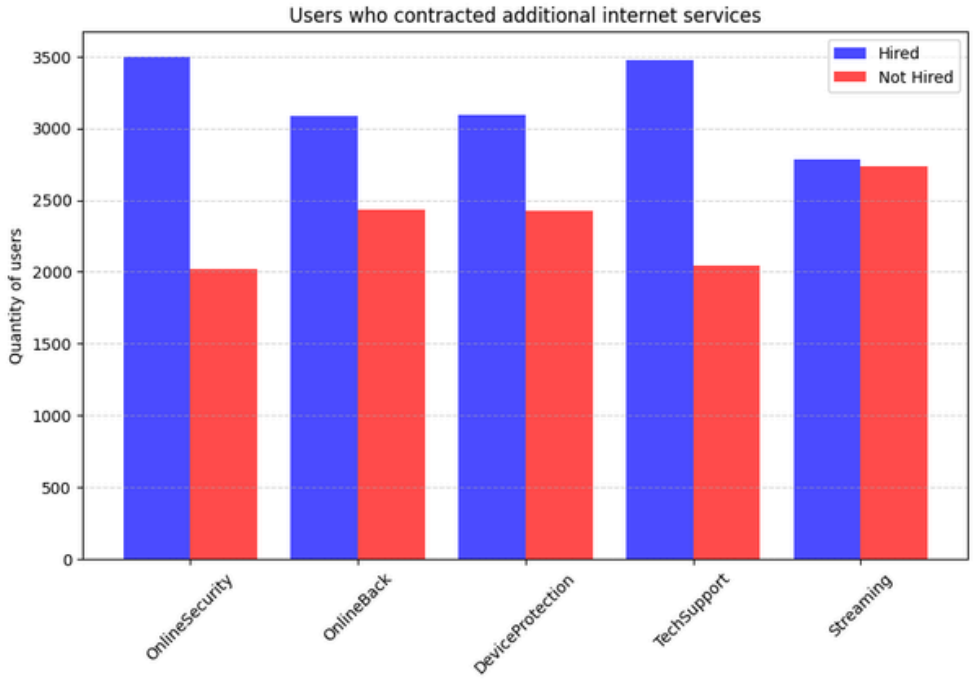
- The year 2014 saw a sharp increase in contracts in February, followed by a steep decline in the following months.
- The year 2019 consistently had the highest number of contracts throughout most months compared to other years.
- Other years, such as 2015–2018, show relatively stable trends with minor fluctuations.
- The contracts in 2020 appear significantly lower, especially in the early months, which aligns with the observed drop in the yearly trend.

General Observations:

- The data suggests that 2019 was a strong year for contracts, while 2020 experienced a sharp decline.
- The early months of the year tend to have higher contract volumes in some years (e.g., 2014 and 2019).
- The consistency of contract numbers from 2015 to 2018 indicates a relatively stable demand in those years.



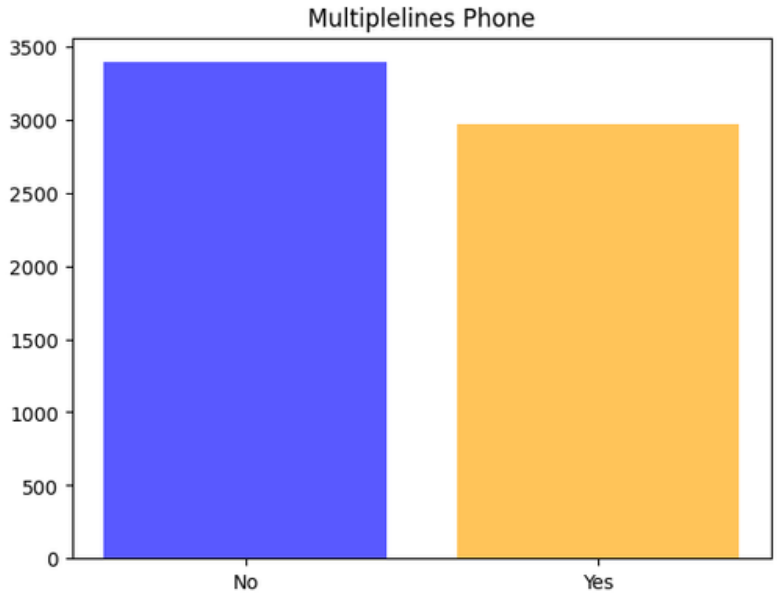
Most people subscribe to fiber optic services.



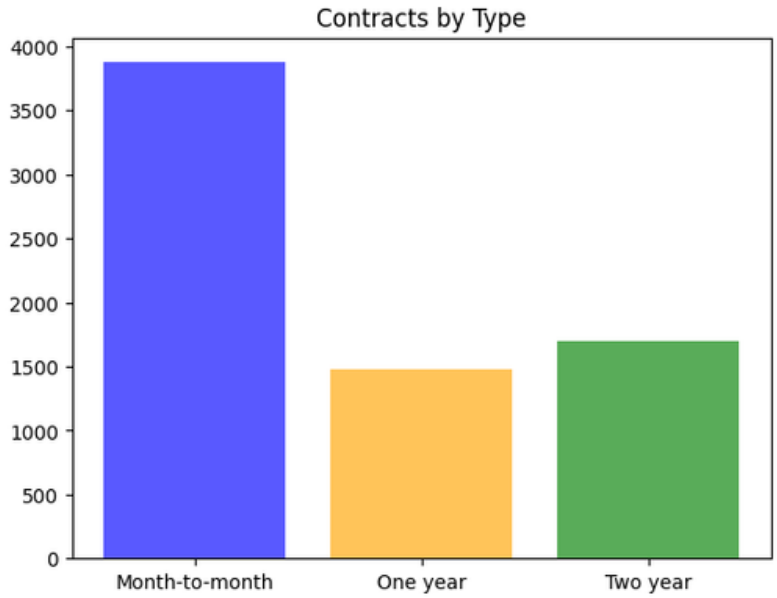
General Trend:
More users tend to subscribe to additional internet services rather than opting out, as indicated by the dominance of blue bars over red ones.

Services:
Tech Support and Online Security appear to be the most frequently hired additional services, as they have the highest number of users subscribed.

Streaming services with nearly equal numbers of users hiring and not hiring.



The number of people who subscribe to multiple lines and those who do not is very similar.

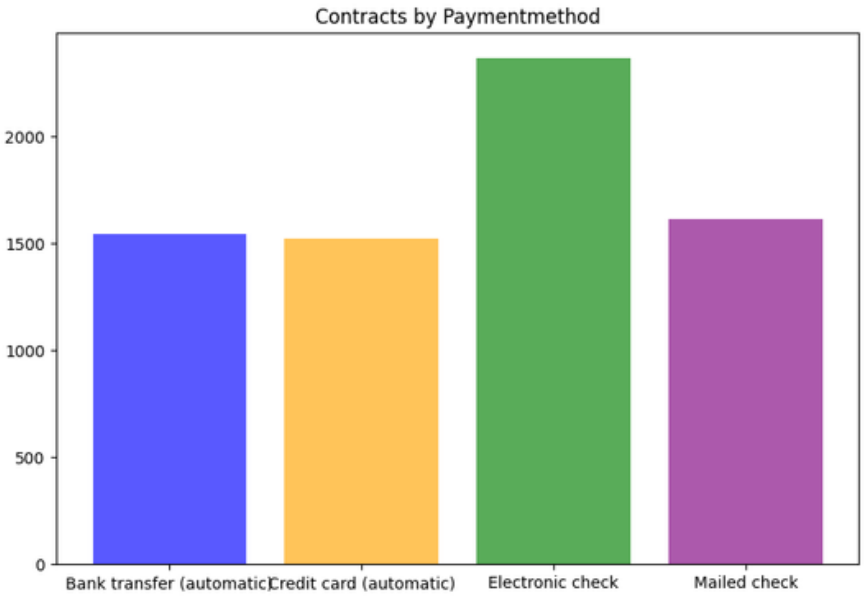


Preference for Month-to-Month Contracts

The majority of customers prefer month-to-month contracts. This suggests that users value flexibility and may avoid long-term commitments.

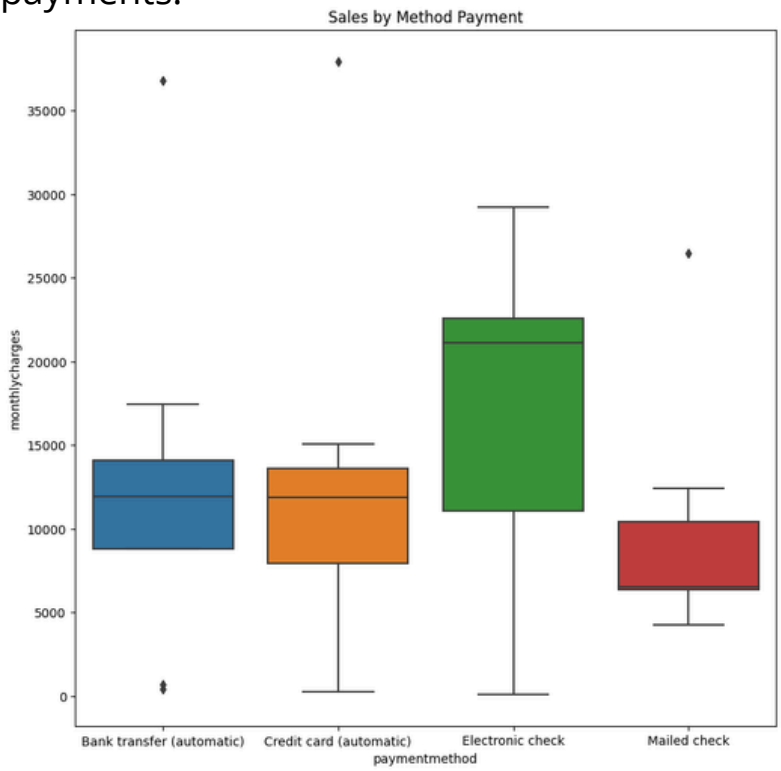
Potential Business Implications

The high number of month-to-month contracts may indicate a higher customer churn rate, as users can easily switch providers.



Preference for Automatic Payments

The most commonly used payment methods are bank transfer (automatic) and credit card (automatic), with nearly the same number of contracts.
Most users prefer automated transactions for convenience and to avoid missed payments.



Higher Revenue from Automatic Payments:

- The bank transfer (automatic) and credit card (automatic) payment methods generate the highest revenues.
- This indicates that customers using automatic payments may have longer retention periods or higher spending habits.

Business Insights:

- Encouraging customers to switch to automatic payment methods could help stabilize revenue and reduce payment delays.
- Strong revenue from automatic payments suggests that subscription-based or recurring payment models are more profitable.

1. Variance Test Results:

- Statistic: 0.6826,
- P-value: 0.5704
- Since the p-value is greater than the significance level (typically 0.05), we cannot reject the null hypothesis.
- This means that the variances between the payment methods are considered equal, justifying the use of ANOVA for comparison.

2. ANOVA Test Results:

- P-value: 0.2864
- Again, the p-value is greater than 0.05, indicating that we cannot reject the null hypothesis.
- This suggests that there is no significant evidence of differences in revenue between the payment methods.

Conclusion:
There is no statistical evidence to suggest that revenue differs significantly between the payment methods (Bank Transfer, Credit Card, Electronic Check, and Mailed Check).

This implies that customers generate similar revenues regardless of their payment method.

4. Predective Data Modeling

Tree different models were created to evaluate and compare performance in predicting the target variable:

Dummy Model (Baseline Model)
A dummy classifier was implemented with the target set to 1, serving as a baseline for comparison.

One-Hot Encoded (OHE) Model with Scaled Data and Upsampling
The dataset was transformed using One-Hot Encoding (OHE) to handle categorical features.
A scaling process was applied to normalize numerical features.
Since the dataset was imbalanced, upsampling was performed to increase the representation of the minority class, ensuring that the model does not favor the majority class.

Ordinal Encoding Model

Ordinal Encoding was applied to categorical variables, assigning numerical labels based on categorical levels.
Upsampling was again performed to balance the dataset.
This approach ensures that the model can handle categorical features efficiently while reducing the dimensionality introduced by One-Hot Encoding.

Dummy Model (Target = 1):

- Dummy Model is a weak benchmark and does not provide valuable predictive power.

One Hot Encoding

- **Logistic Regression:** performs well but does not outperform the Decision Tree Classifier in F1-score or ROC-AUC.
- **Decision Tree Classifier:** with depth 7 is the best model overall, providing high generalization performance.
- **Random Forest:** models tend to overfit, achieving nearly perfect training accuracy, which is not ideal for real-world applications.

Ordinal Encoding

- Tree Classifier is the simplest but weakest in terms of predictive power.
- Random Forest improves performance but suffers from overfitting.
- LightGBM is superior to both in terms of ROC-AUC and generalization, making it the best choice if interpretability is not a major concern.

The best model is CatBoost, as it outperforms all other models in ROC-AUC, F1-Score, and Log Loss. It provides the most balanced and robust performance across all datasets.

Cat Boost Values:

- AUC-ROC Train: 0.9505
- AUC-ROC Valid: 0.8917
- AUC-ROC Test: 0.8860

- Accuracy Train: 0.8756
- Accuracy Valid: 0.8360
- Accuracy Test: 0.8324

- F1-Score Valid: 0.6841
- Log Loss Valid: 0.3494

INTERCONNECT

Final Conclusions

By leveraging predictive modeling and customer insights, Interconnect can proactively reduce churn by targeting at-risk users with personalized offers. The combination of contract flexibility, service bundling, and predictive analytics will help increase customer retention and maximize revenue.