Individual Assignment

SOFTWARE PROGRAMMING FOR BUSINESS & ECONOMICS

Haram Kang
Applied Statistics
20212049

I utilized Python through **Google Colab** for the convenience of this assignment.

Q1. Analysis of telecom_churn.csv

1. Initial Setup

1-1. Handling Missing and Outlier Values

Confirmed the absence of missing values through code.
Extracted basic statistics and confirmed the absence of outliers.

2. Useful Information for Understanding the Telecom Market

2-1. Overview of Customer Churn

The primary interest for our boss would be the customer churn rate. By examining the 'Churn' variable, it was found that 85.5% of the data points are labeled 'False' (not churned) and 14.5% are labeled 'True' (churned). This indicates an imbalanced dataset.

```
selected_columns = ['Number vmail messages', 'Total day minutes', 'Total day calls', 'Total day charge', 'Total eve minutes', 'Total eve calls', 'Total eve charge', 'Total night minutes', 'Total night calls', 'Total night charge', 'Total intl minutes',
tel[selected_columns].describe()
```

	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
std	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
50%	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
75%	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
max	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

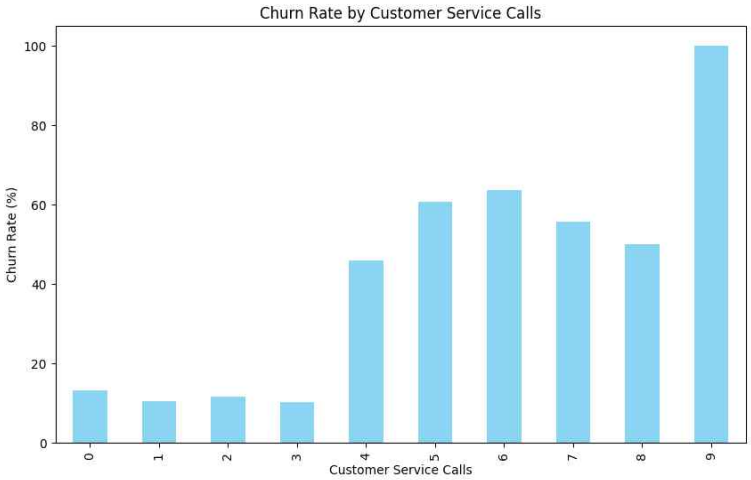
To identify the variables most related to churn, the 'Churn' variable was converted to numerical format ('True' as 1 and 'False' as 0). Correlation analysis was performed on the numeric columns, and it was found that the variables with the highest correlation to churn are:

Customer service calls: 0.208750
Total day minutes: 0.205151
Total day charge: 0.205151

These three variables were further analyzed and visualized to understand their impact on churn rates.

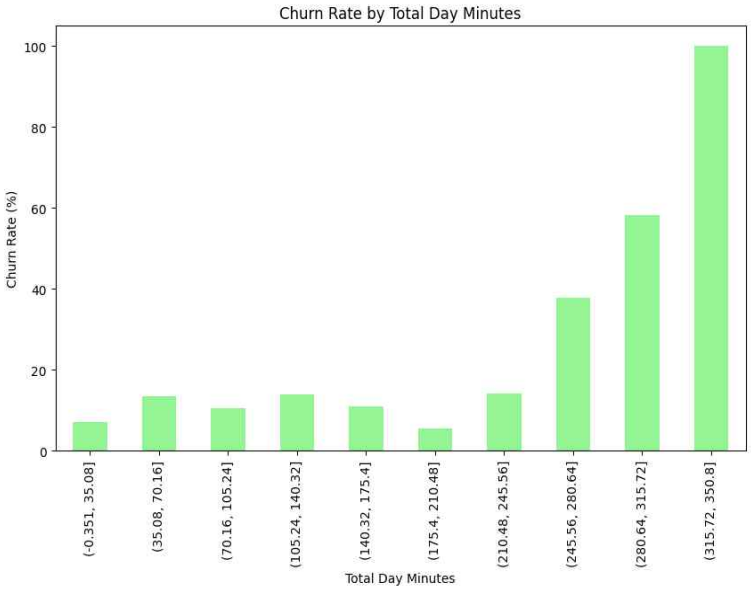
2-2. Visualization of Key Variables

1. Customer Service Calls vs. Churn Rate:



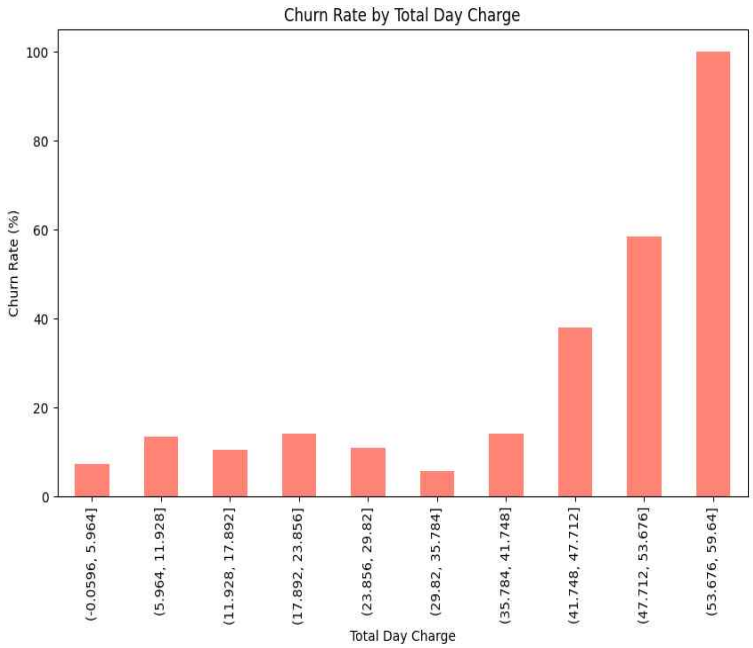
Customers who make more customer service calls tend to have a higher churn rate.

2. Total Day Minutes vs. Churn Rate:



Higher total day minutes usage correlates with higher churn rates.

3. Total Day Charge vs. Churn Rate:



Similar to total day minutes, higher total day charges are associated with higher churn rates.

These figures are of particular interest to our boss as they highlight the factors most strongly associated with customer churn. Understanding these relationships can help in developing targeted strategies to reduce churn.

3. Predictive Modeling for Customer Churn

3-1. Data split

The dataset was split into 80% training data and 20% test data. Three different models were trained to predict customer churn: Logistic Regression, Random Forest, and XGBoost.

3-2. Model Evaluation

Due to the imbalanced nature of the dataset, both accuracy and F1-score were used to evaluate the models. The results are as follows:

- Logistic Regression:
Accuracy: 0.8546
F1-score: 0.2595

- Random Forest:
Accuracy: 0.9325

F1-score: 0.7239

- XGBoost:
Accuracy: 0.9610
F1-score: 0.8571

XGBoost showed the highest scores in both accuracy and F1-score, making it the best model for this task. This model's superior performance can be attributed to its ability to handle complex relationships and interactions between variables more effectively than the other models.

4. Strategic Implications for Telecom Companies

Based on the results from parts (a) and (b), the following strategic implications can be suggested for telecom companies:

1. Improve Customer Service:
High correlation between customer service calls and churn rate suggests that improving customer service could significantly reduce churn. Investing in better training for customer service representatives and reducing wait times could be beneficial.

2. Monitor High Usage Customers:
Customers with high total day minutes and charges are more likely to churn. Implementing loyalty programs or offering personalized plans for these high-usage customers might help in retaining them.

3. Proactive Engagement:
Use predictive models like XGBoost to identify customers at high risk of churn and engage with them proactively. This could involve targeted marketing campaigns, special offers, or check-in calls to address any potential issues.

4. Analyze and Address Underlying Issues:
Conduct further analysis to understand why high usage correlates with churn. It may be due to billing issues, network problems, or other service quality issues that need to be addressed.

By focusing on these areas, telecom companies can better manage customer relationships and reduce churn rates, ultimately improving their market position and profitability.

Q2. food mart data.csv

1. Initial Setup

1-1. Handling Missing and Outlier Values

Upon examining the dataset using is.null, we confirmed that there are no missing values in the Food Mart data. Additionally, by comparing the 75th percentile and the maximum values in the summary statistics, we determined that there are no outliers in the dataset.

2. Please provide any useful information (e.g., figures, tables, etc.) to understand customers and product categories.

2-1. Feature Engineering and Derived Metrics

To provide deeper insights into customers and product categories, we performed extensive feature engineering. The following derived metrics were created:

Income: This metric was calculated as the difference between total sales and cost. Specifically, Income = (store_sales(in millions) * 1,000,000) - cost, which was then normalized back to millions for consistency.

Income Ratio: This metric represents the ratio of income to total sales, calculated as Income Ratio = Income / store_sales(in millions). This ratio provides insight into the profitability relative to sales.

Unit Price: The average price per product unit, calculated as Unit Price = store_sales(in millions) / unit_sales(in millions). This metric helps understand the pricing strategy for different products.

Sales per Square Foot: This metric evaluates the sales efficiency relative to the store size, calculated as Sales per Square Foot = store_sales(in millions) / Store_sqft. It is useful for assessing how effectively the store space is utilized to generate sales.

Service Density: This metric measures the density of various services offered within the store, calculated as Service Density = (Coffee_bar + Video_store +

Salad_bar + Prepared_food + Florist) / Store_sqft. This helps in understanding the concentration of services available to customers within the store.

2-2. Correlation Matrix Analysis

We extracted the correlation matrix for all features to understand the relationships between different variables. The correlation matrix revealed several key insights:

- Various metrics derived from sales, such as income and income ratio, showed significant correlations with each other, which is expected given their direct derivation from sales figures.

- Apart from these expected correlations, there were no strong correlations between other features, indicating that most features are relatively independent of each other.

This analysis suggests that while sales-related metrics are closely intertwined, other variables such as customer demographics and store attributes do not exhibit strong linear relationships with each other. This independence can be advantageous in predictive modeling, as it indicates a diverse set of features with unique contributions to the target variable.

By understanding these derived metrics and their relationships, we can gain valuable insights into customer behavior and product category performance, aiding in strategic decision-making for the Food Mart.

3. Please split the data to a training dataset and a test dataset, and build a regression model to understand important components of supermarket chains and predict the cost of media campaigns and sales in the food marts on the basis of the features provided. Finally, please justify your model and variables (compared to other models).

3-1. Data Splitting and Model Building

The dataset was split into training and test sets with an 80-20 ratio. Two regression models were employed: Linear Regression and XGBoost Regression.

3-2. Results

The results from the two regression models are as follows:

1. Linear Regression:
Mean Squared Error (MSE): 665.4760423308475
R-squared (R2): 0.25939793658877874
2. XGBoost Regression:
Mean Squared Error (MSE): 2.1670256850423675
R-squared (R2): 0.9975883373829864

3-3. Analysis and Justification

The XGBoost Regression model demonstrated significantly better performance compared to the Linear Regression model. The XGBoost model achieved a much lower Mean Squared Error and a substantially higher R-squared value, indicating a more accurate and reliable model for predicting the cost of media campaigns and sales in food marts.

The R-squared value for the XGBoost model suggests that it explains approximately 99.76% of the variance in the target variable, which is a strong indication of its effectiveness in capturing the complex relationships within the data. In contrast, the Linear Regression model's R-squared value of 0.259 suggests it explains only about 25.94% of the variance, making it a less suitable choice for this task.

Given the superior predictive performance of the XGBoost model, as evidenced by the significantly higher R-squared value and lower Mean Squared Error, it is justified to choose the XGBoost Regression model for this analysis. This model effectively captures the intricate patterns in the data, making it the most appropriate choice for predicting the cost of media campaigns and sales in supermarket chains.

3-4. Exploration of Additional Regression Models Using PyCaret

In order to explore a wider range of regression models, I employed PyCaret, an open-source, low-code machine learning library in Python that automates machine learning workflows. This workflow enabled me to train various models and evaluate their performance based on Mean Squared Error (MSE).

-PyCaret Workflow and Results

Utilizing PyCaret, I trained multiple regression models with appropriate preprocessing applied to each feature. The data was divided into 10 folds for cross-validation. The results of the best 3 training are summarized below:

1. Linear Regression

MSE: 0.1751

MAE: 0.0610
R2: 0.9999

2. Extra Trees Regressor
MSE: 0.1574
MAE: 0.1082
R2: 0.9999

3. Extreme Gradient Boosting (XGBoost)
MSE: 0.9336
MAE: 2.1836
R2: 1.9976

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
lr	Linear Regression	0.1751	0.0610	0.2470	0.9999	0.0028	0.0019
et	Extra Trees Regressor	0.1574	0.1082	0.3284	0.9999	0.0036	0.0016
xgboost	Extreme Gradient Boosting	0.9336	2.1836	1.4770	0.9976	0.0150	0.0099
rf	Random Forest Regressor	0.9506	2.5838	1.6073	0.9971	0.0169	0.0101
lightgbm	Light Gradient Boosting Machine	1.1778	2.8663	1.6926	0.9968	0.0170	0.0123
dt	Decision Tree Regressor	1.8647	19.9344	4.4633	0.9777	0.0466	0.0197
gbr	Gradient Boosting Regressor	23.8958	782.3236	27.9699	0.1252	0.2995	0.2762
ada	AdaBoost Regressor	25.0863	844.5820	29.0615	0.0555	0.3110	0.2908
ridge	Ridge Regression	25.5473	876.9790	29.6137	0.0193	0.3164	0.2963
br	Bayesian Ridge	25.5500	877.1645	29.6168	0.0191	0.3165	0.2963
lasso	Lasso Regression	25.7047	883.8105	29.7288	0.0117	0.3179	0.2985
llar	Lasso Least Angle Regression	25.7047	883.8105	29.7288	0.0117	0.3179	0.2985
en	Elastic Net	25.7132	884.3226	29.7374	0.0111	0.3180	0.2986

-Analysis and Conclusion
Interestingly, Linear Regression, which showed poor performance in previous analyses, demonstrated significantly improved results when processed through PyCaret. This suggests that PyCaret's preprocessing capabilities may have effectively enhanced the model's predictive accuracy by optimizing feature handling and model training.

4. Given the results from a-b), please develop any customer or retail chain strategies for the supermarkets to acquire new customers and boosts sales.

Based on the analysis, although there appeared to be no simple correlations

between individual variables (features), machine learning models successfully predicted the cost accurately. This finding suggests that leveraging these models could enable effective cost optimization strategies for the stores.

By utilizing machine learning regression models, despite the absence of straightforward correlations between features, we achieved precise predictions of the cost incurred in acquiring customers. This capability opens avenues for implementing robust cost optimization strategies within the store operations.

Understanding that traditional linear correlations may not fully capture the intricate relationships within the dataset, the machine learning approach proves invaluable in forecasting costs with high accuracy. This insight empowers supermarkets to make informed decisions aimed at optimizing operational expenses and enhancing profitability.

In conclusion, the application of advanced machine learning techniques not only improves predictive accuracy but also provides actionable insights for optimizing costs and fostering sustainable growth in competitive retail environments.