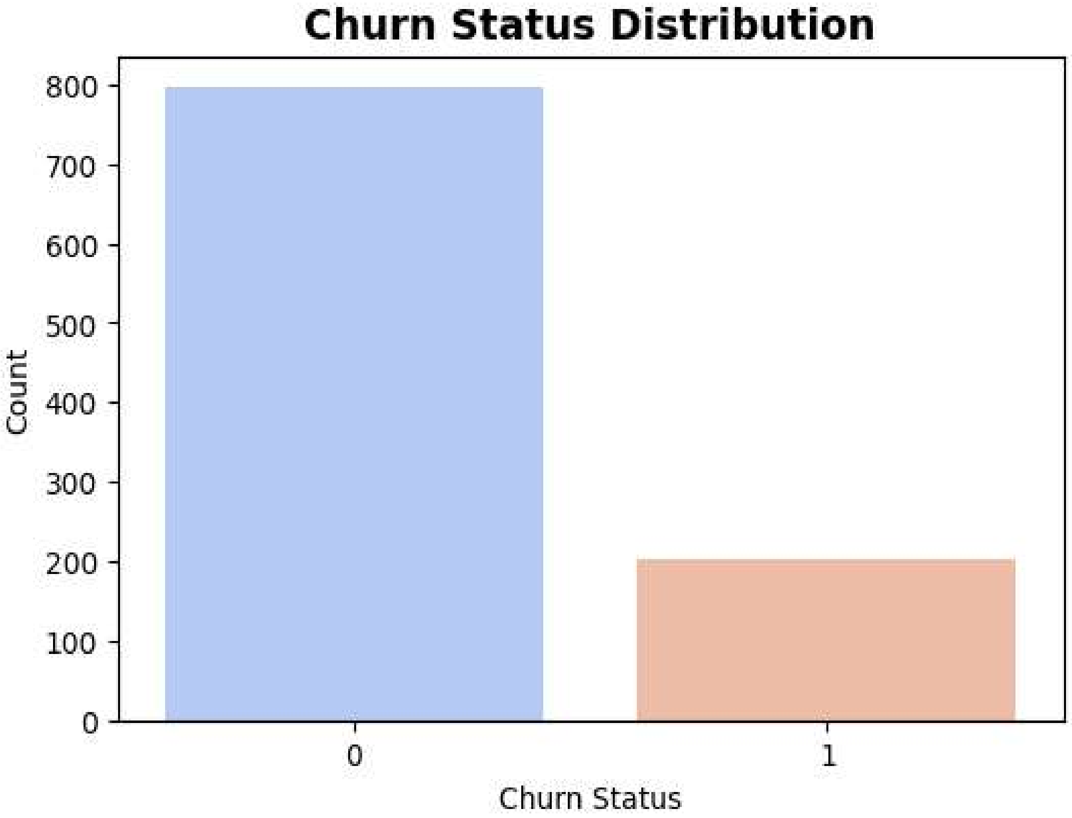
Customer Churn Prediction – Lloyds Banking Group Data Science Internship

As part of my data science internship with Lloyds Banking Group, I worked on an end-to-end machine learning project to predict **customer churn** using real-world banking data. The goal was to analyze customer behavior, perform exploratory data analysis (EDA), and build predictive models that identify customers likely to discontinue services.

1. Exploratory Data Analysis (EDA)

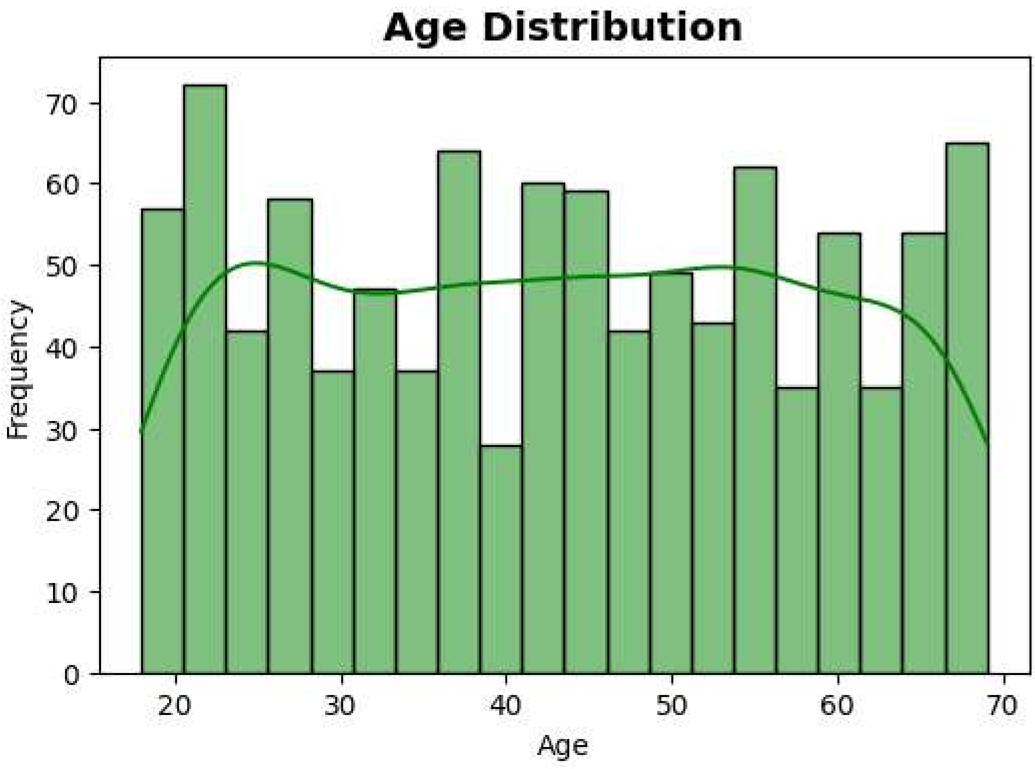
I began by cleaning and preparing the data, handling missing values, and performing one-hot encoding (OHE) on categorical variables. EDA was conducted to gain deeper insights into factors influencing churn. I created several visualizations, including:

Churn distribution: Identified class imbalance, with a significantly lower proportion of churned customers.



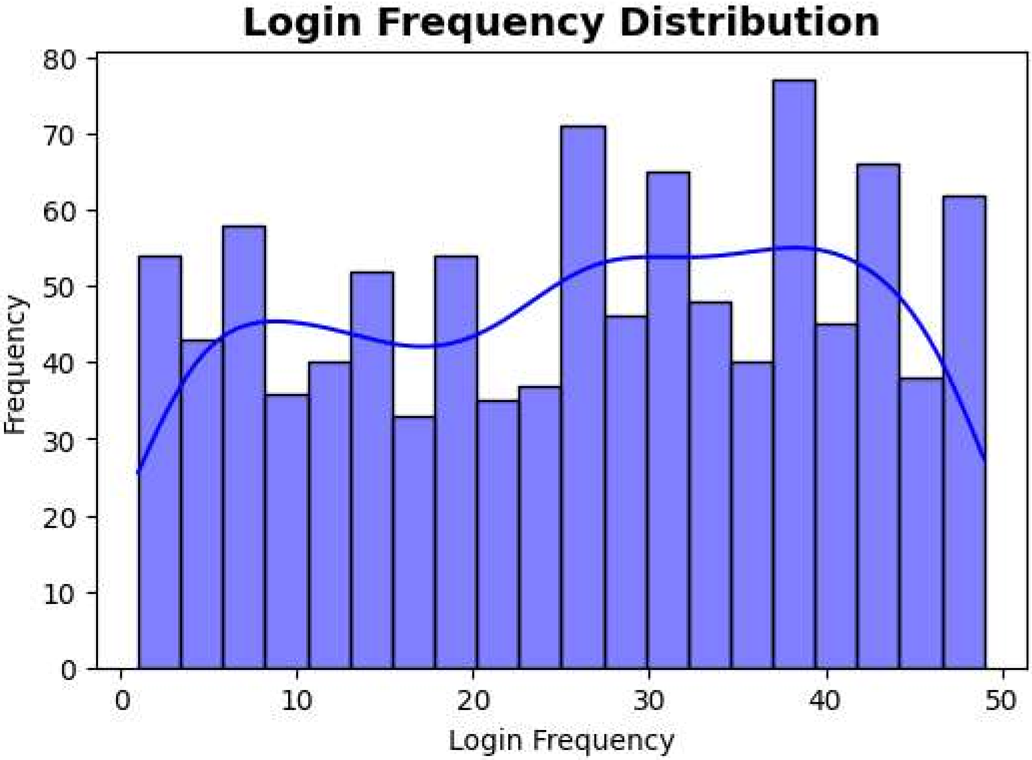
the dataset represented in the chart shows a heavily imbalanced distribution of churn status, with a large majority of observations indicating "no churn" compared to a relatively small number indicating "churn."

Age distribution: Highlighted age groups with higher churn risk



This image is a histogram that visually represents the age distribution within a given dataset. The horizontal axis, labeled "Age," displays various age ranges, spanning from roughly 15-20 years up to 70 years. The vertical axis, labeled "Frequency," indicates the count or number of individuals that fall into each specific age group. The green bars illustrate these frequencies, with their height directly corresponding to the number of individuals in that particular age bracket. Superimposed on these bars is a smooth green curve, which serves as a Kernel Density Estimate (KDE). This curve provides a continuous and smoothed representation of the underlying probability distribution of ages, offering a clearer picture of the overall shape and density of the data than the discrete bars alone. Observing both the bars and the curve reveals that the age distribution is not uniform; there are noticeable variations in the prevalence of different age groups, suggesting that certain age ranges are more common within the dataset than others. The curve generally indicates a higher concentration of individuals in younger to middle-aged groups, with a potential peak in the early 20s, followed by a relatively stable period, and then a decline in the older age categories.

Login frequency distribution: Showed a strong link between low login activity and churn.



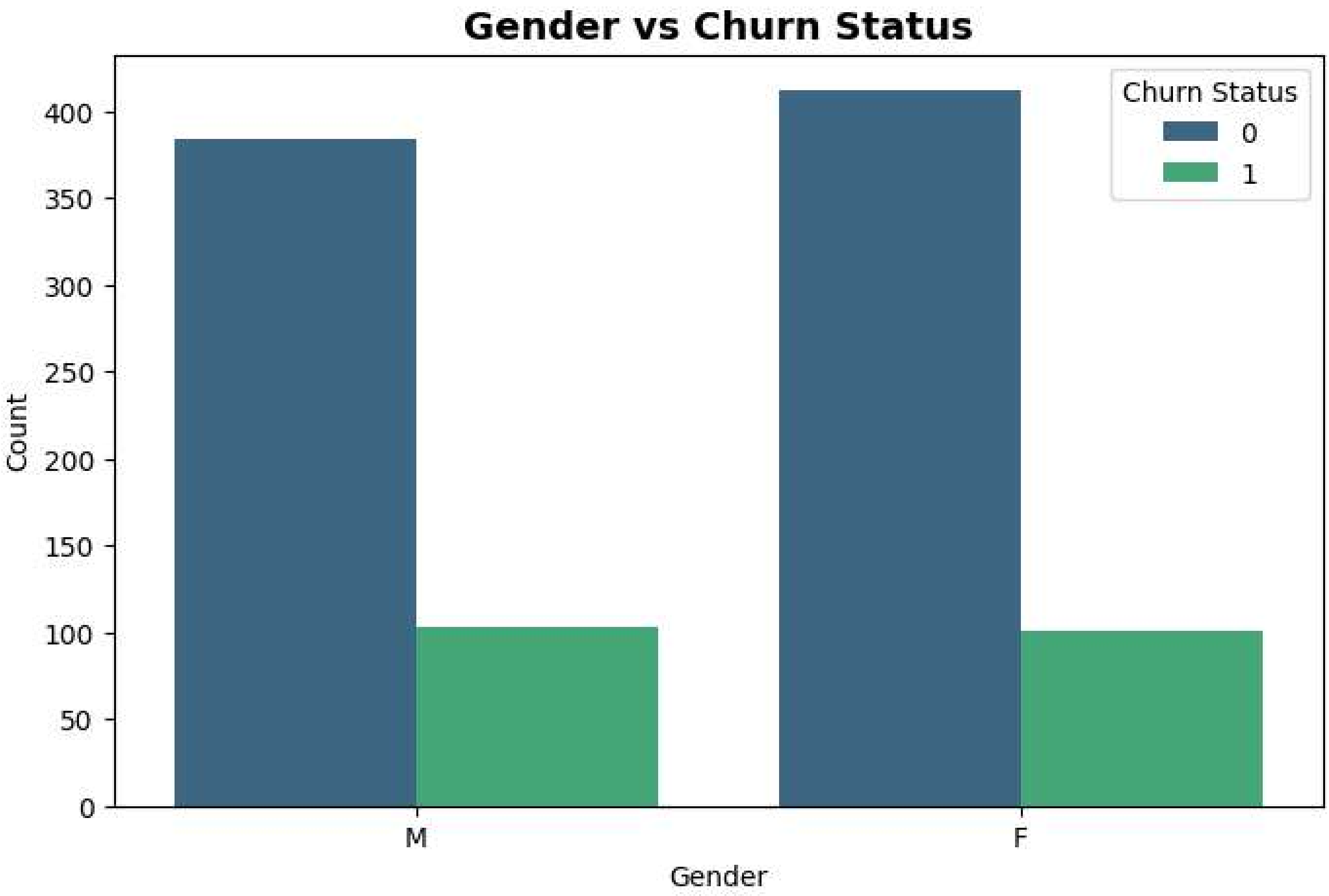
This image presents a histogram titled "Login Frequency Distribution," which illustrates the distribution pattern of how often users log in.

The horizontal axis ("Login Frequency") ranges from 0 to 50, representing different counts of logins. The vertical axis ("Frequency") indicates the number of individuals or observations that fall within each corresponding login frequency range. The blue bars show the raw counts for each bin, with their varying heights reflecting different frequencies. For instance, there's a noticeable peak around a login frequency of 40, where the bar approaches a frequency of 80. Conversely, there are dips, such as around a login frequency of 20, where the frequency is significantly lower.

Superimposed on the bars is a smooth blue curve. This curve is a Kernel Density Estimate (KDE), providing a continuous and smoothed representation of the underlying distribution. It helps in visualizing the overall trend and density more clearly than the discrete bars alone. From the curve, we can infer that the frequency generally rises at lower login counts, dips slightly around the 15-20 mark, and then shows a more pronounced peak in the range of roughly 25 to 45 login frequencies. This indicates that a significant portion of users fall into these higher login frequency categories. The frequency then declines again as login counts approach 50.

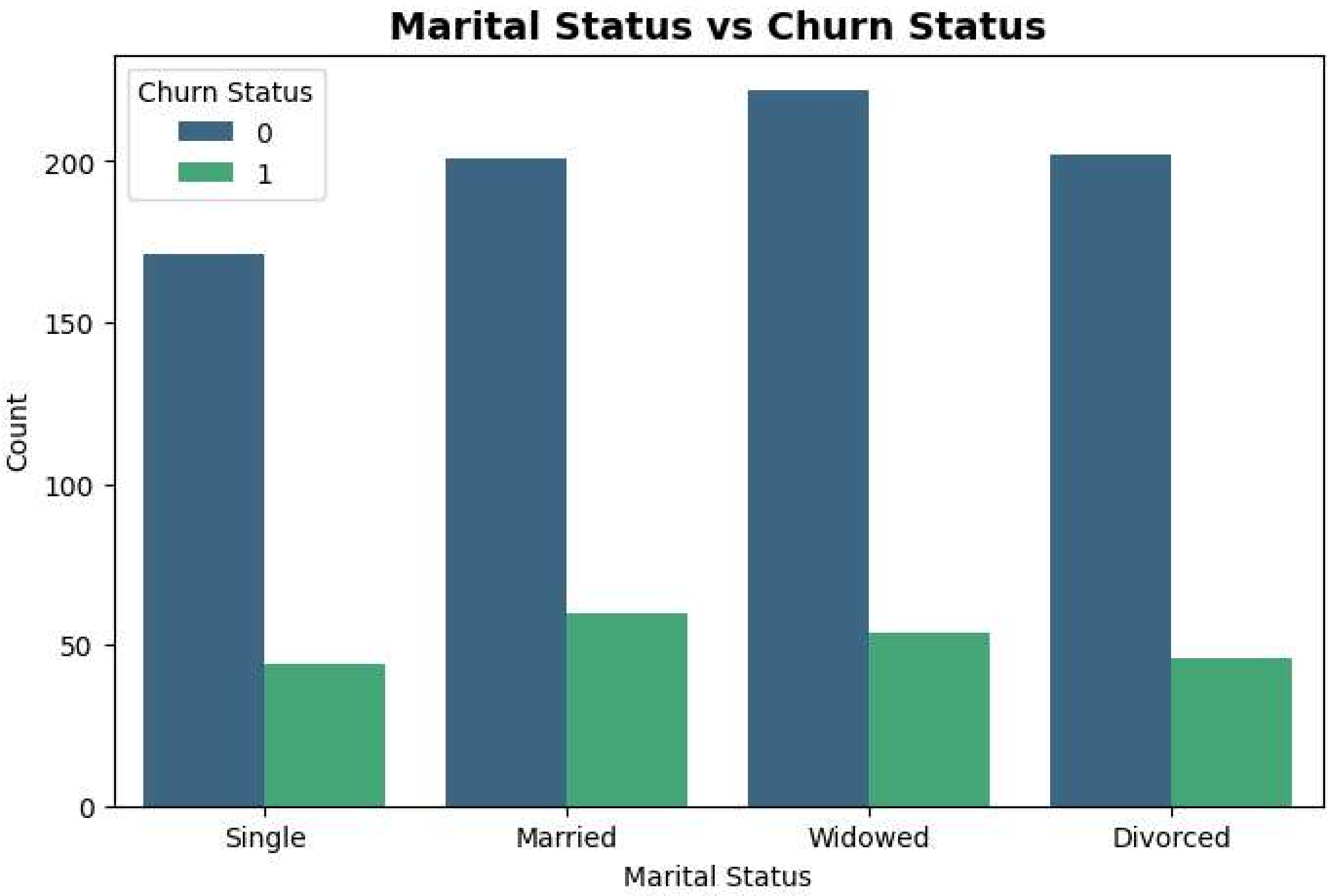
In summary, the chart reveals that login frequencies are not uniformly distributed. Instead, there are specific ranges where login activity is more concentrated, particularly in the mid-to-high frequency range, suggesting that a good number of users are quite active

Gender vs. Churn Status: Minor differences observed between male and female churn rates.



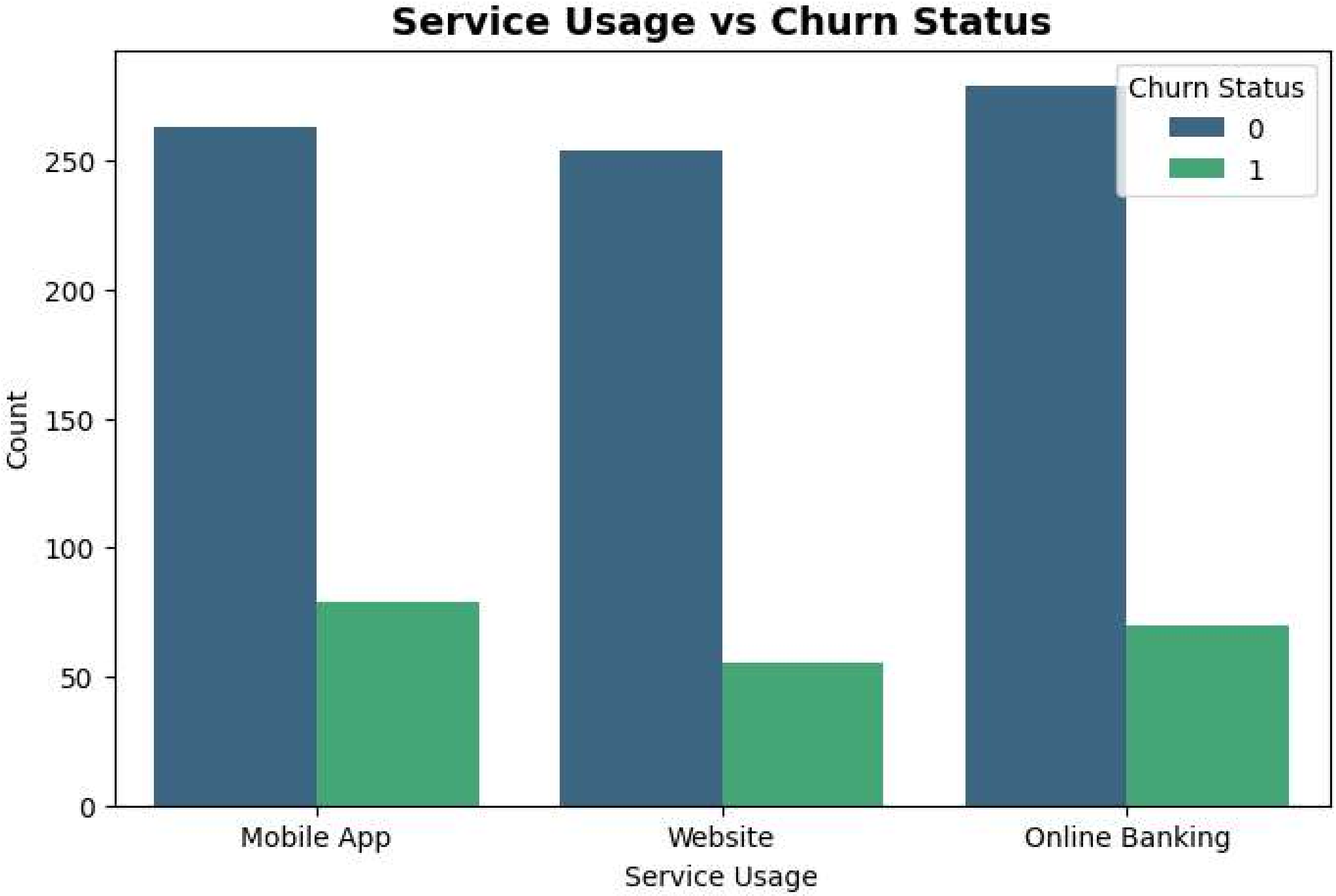
The "Gender vs Churn Status" chart is a clustered bar chart that illustrates the relationship between a user's gender (Male or Female) and their churn status (0 for no churn, 1 for churn). For both male ('M') and female ('F') categories, the blue bars representing 'Churn Status 0' (no churn) are significantly taller than the green bars representing 'Churn Status 1' (churn). Specifically, there are approximately 380 males and 410 females who have not churned, while about 100 males and 100 females have churned. This visual comparison suggests that, within this dataset, the proportion of users who churn appears to be relatively consistent across both male and female genders, indicating that gender does not seem to be a primary differentiating factor in churn behavior.

Marital Status vs. Churn Status: Certain marital groups exhibited higher churn tendencies.



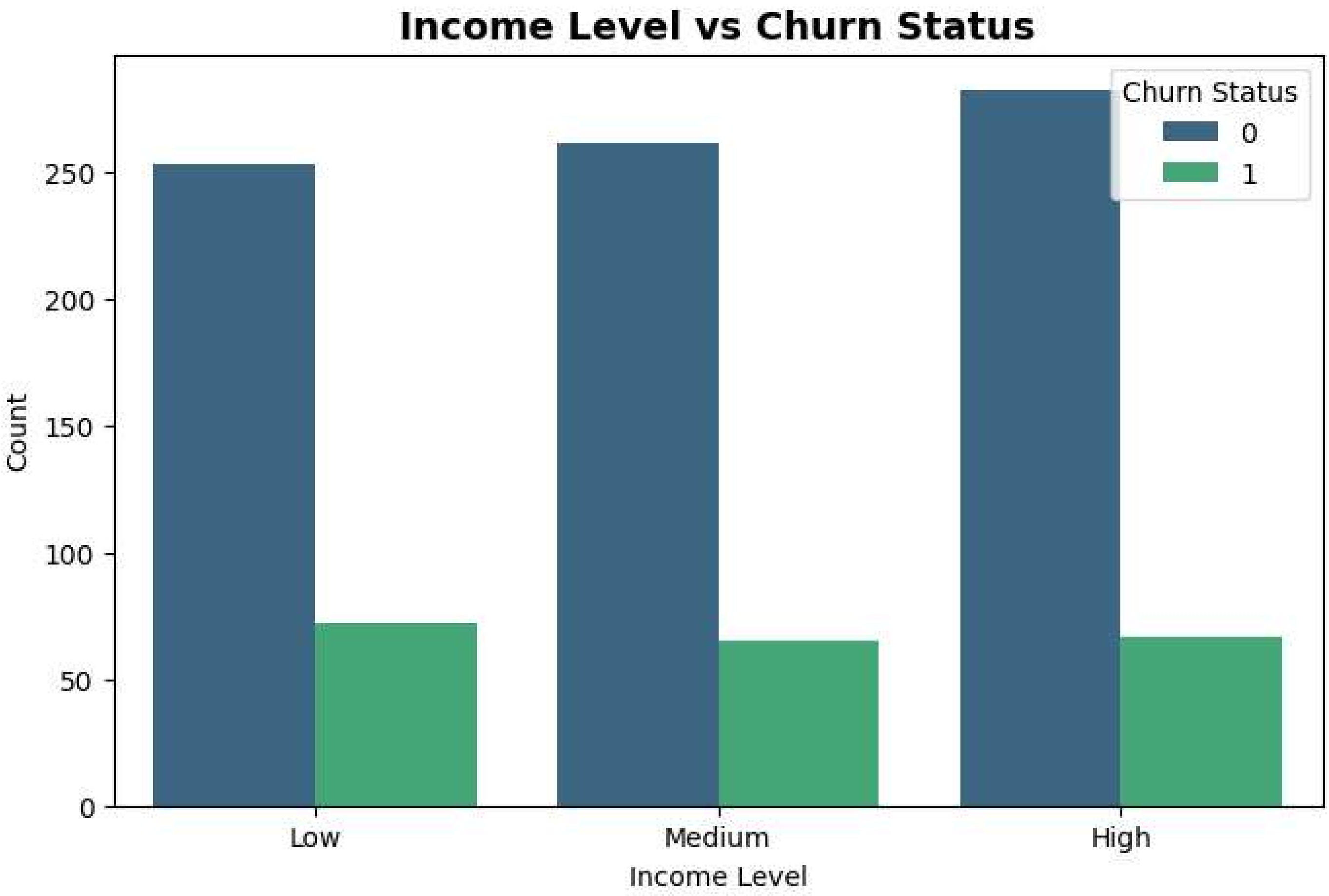
The "Marital Status vs Churn Status" chart is a clustered bar chart that illustrates the relationship between different marital statuses (Single, Married, Widowed, Divorced) and churn status (0 for no churn, 1 for churn). For every marital status displayed, the count of individuals who did not churn (represented by the darker blue bars with 'Churn Status 0') is considerably higher than the count of those who did churn (represented by the green bars with 'Churn Status 1'). While there are variations in the absolute number of individuals within each marital status, for instance, a higher count of non-churned 'Widowed' individuals compared to 'Single' individuals, the proportion of churned individuals relative to non-churned individuals appears to be consistently small across all marital statuses. This observation suggests that marital status, on its own, may not be a significant determining factor or a strong predictor of customer churn in the analyzed dataset.

Service Usage vs. Churn Status: High service engagement correlated with lower churn.



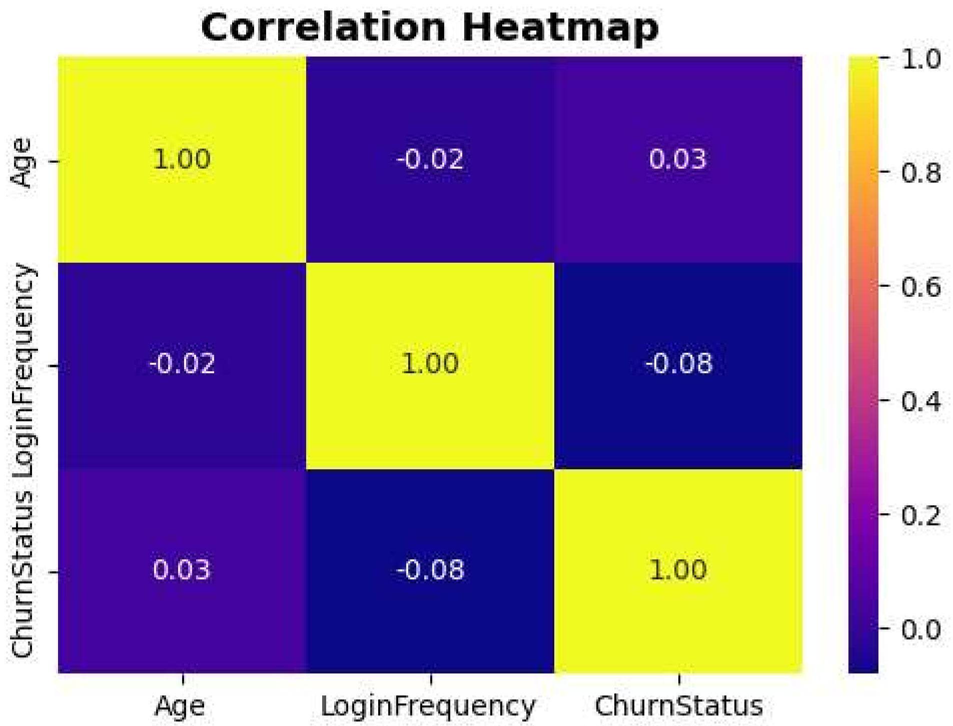
The chart visually demonstrates a strong correlation between high service engagement and lower customer churn. Across all three presented service usage categories—Mobile App, Website, and Online Banking—the number of customers who have not churned significantly outnumbers those who have churned. For instance, roughly 260 non-churned customers use the Mobile App compared to about 80 churned customers. Similarly, approximately 250 non-churned customers use the Website, whereas around 55 churned customers do. Online Banking shows the highest engagement with about 275 non-churned users versus roughly 70 churned users. This consistent pattern across all services clearly indicates that customers who are more engaged with the provided services are less likely to discontinue their relationship.

Income Level vs. Churn Status: Lower income levels showed increased churn risk.



This bar chart, titled "Income Level vs Churn Status," illustrates the relationship between customer income levels and churn. It shows that in all three income categories—Low, Medium, and High—the number of customers who have not churned significantly exceeds those who have churned. While the absolute number of churned customers remains relatively consistent across all income levels (around 65-70), the overall trend indicates a higher proportion of non-churned customers as income level increases. This aligns with the chart's guiding statement that "Lower Income levels showed increased churn risk," suggesting that while retention is strong across the board, the relative likelihood of churn is slightly higher for customers in lower income brackets.

Correlation heatmap: Provided an overview of relationships among numerical features.



This image presents a Correlation Heatmap, which visually summarizes the relationships between Age, LoginFrequency, and ChurnStatus. The color intensity ranges from dark purple for negative correlations to bright yellow for positive correlations. As expected, each variable is perfectly correlated with itself, shown by the 1.00 values on the diagonal. The map reveals a very weak negative correlation of -0.02 between Age and LoginFrequency, suggesting a negligible inverse relationship. Age and ChurnStatus exhibit a very weak positive correlation of 0.03, indicating an almost non-existent tendency for churn to increase with age. The strongest, though still weak, relationship is a negative correlation of -0.08 between LoginFrequency and ChurnStatus, implying that more frequent logins might be marginally associated with a lower likelihood of churning. In essence, the heatmap demonstrates that there are no strong linear relationships among these particular numerical features.

2. Data Preprocessing

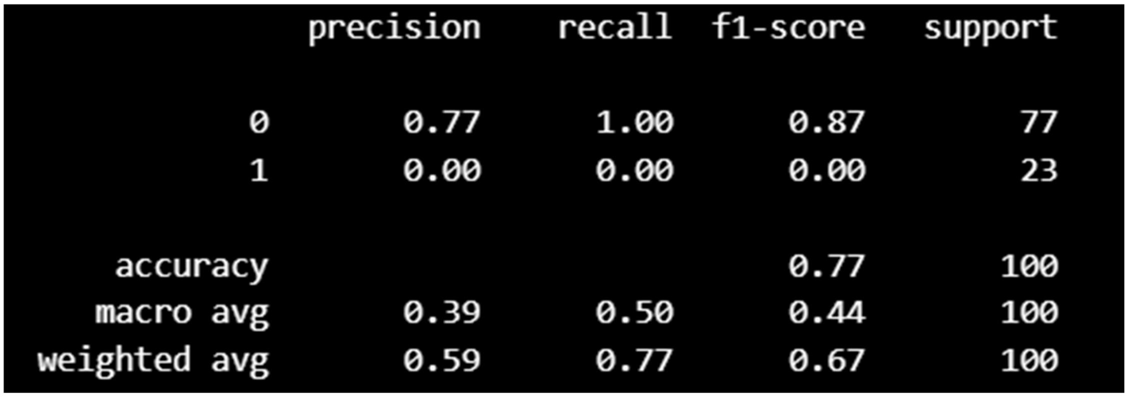
- Applied **one-hot encoding** for categorical features.
- Standardized numerical variables for model compatibility.
- Addressed class imbalance through careful evaluation of metrics beyond simple accuracy.

3. Model Building and Evaluation

I experimented with several **classification algorithms**, including:

- **Logistic Regression**
- **Support Vector Machine (SVM) Classifier**
- **Naive Bayes**

Initial models achieved approximately **77% accuracy**. I then performed **hyperparameter tuning** to optimize performance and evaluated models using the **classification report**, focusing on precision, recall, and F1-score to account for data imbalance.



```
              precision    recall  f1-score   support

           0       0.77      1.00      0.87        77
           1       0.00      0.00      0.00        23

    accuracy                           0.77       100
   macro avg       0.39      0.50      0.44       100
weighted avg       0.59      0.77      0.67       100
```

4. Outcome

The analysis demonstrated that customer engagement features, income level, and service usage were strong indicators of churn. The models provided a reliable baseline for predicting churn, and the insights gained can support retention strategies, such as targeted outreach and personalized customer engagement.

5. Conclusion

This project enhanced my hands-on experience in data preparation, EDA, and machine learning model development for real-world business problems. The ability to identify at-risk customers can directly contribute to reducing churn rates, improving customer satisfaction, and increasing overall business value for Lloyds Banking Group.