Predictive Modeling for Customer Churn

Jibin Sebastian

23-04-2024

Objective: Build a predictive model to identify customers who are likely to churn.

Data Collection and Preprocessing:

Code: src\data_collection_preprocessing.py

This script is designed to preprocess and split a customer churn dataset into training (80%), validation (10%), and test (10%) sets as a best practice for machine learning.

Feature engineering: As part of feature engineering, generated three extra features, for instance income level, Age group and Loyalty level of customers to understand more about data and business scenarios.

Data Exploration: **Observation and findings**

Code: src\explanatory_data_analysis.py

This script provides functions to visualize and analyse a customer churn train dataset.

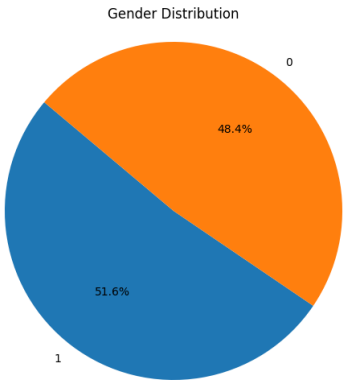
Explanatory Data Analysis on Train dataset

Here are some insights from the summary statistics (data\summary_statistics.csv) of the customer churn train dataset:

1. Age Distribution:

- The average age of the customers is approximately 44 years, with a standard deviation of 14.71. The ages range from 18 to 69 years. The majority of the customers fall into the 'Adult' age group (59.25%), followed by 'Senior' age group (17.25%).

2. Gender Distribution:



- About 51.63% of the customers are male (Gender=1) and 48.38% are female (Gender=0).

3. Annual Income:

- The average annual income is approximately $85,300, with a standard deviation of $38,526.

- Income ranges from $20,077 to $149,972.

- Most customers have an income level categorized as 'High', as 73.5% fall into this category.

4. Years with Company:

- On average, customers have been with the company for approximately 10 years, with a standard deviation of 5.51 years.

- The tenure ranges from 1 to 19 years.

- Most customers are categorized under 'High' loyalty level, indicating that they have been with the company for a longer duration (73.5%).

5. Number of Support Calls:

- The average number of support calls made by customers is approximately 4.43, with a standard deviation of 2.84.

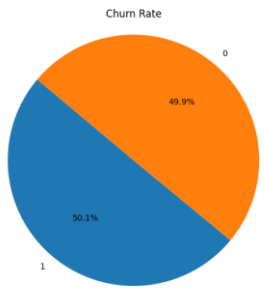
- The number of support calls ranges from 0 to 9.

6. Monthly Charges:

- The average monthly charge is approximately $112.36, with a standard deviation of $50.90.

- Monthly charges range from $20 to $199.

7. Churn Rate:



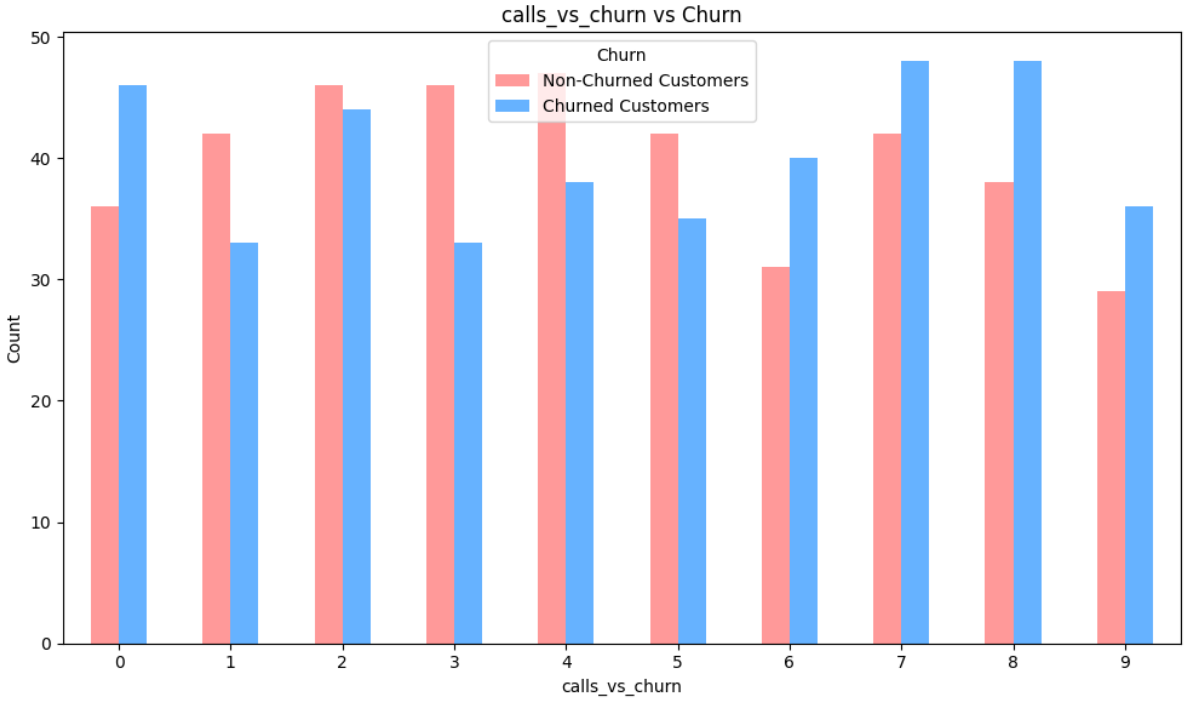
- The average churn rate is 50.1%, which means that the dataset is relatively balanced between churned and non-churned customers.

Summary statics for income and Years with company feature in train dataset

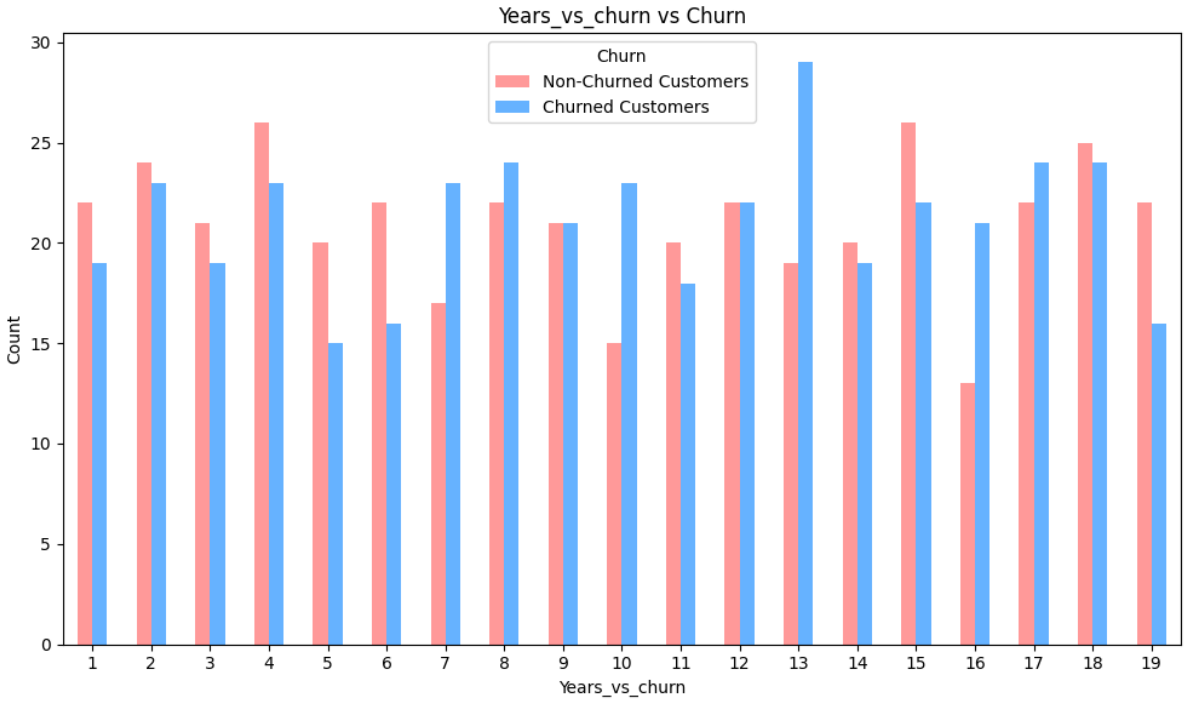
	Income	Years_with_Company
count	800.000000	800.000000
mean	85300.180000	10.032500
std	38526.297736	5.512888
min	20077.000000	1.000000
25%	51464.750000	5.000000
50%	84963.500000	10.000000
75%	119835.250000	15.000000
max	149972.000000	19.000000

A few insights:

Income:
The majority of customers have an income below the mean ($85,300), as indicated by the 25th and 50th percentiles.
There is a significant income range, with the lowest income at $20,077 and the highest at $149,972.
Years with Company:
The average customer has been with the company for about 10 years.
The majority of customers have been with the company for 15 years or less, as shown by the 75th percentile.



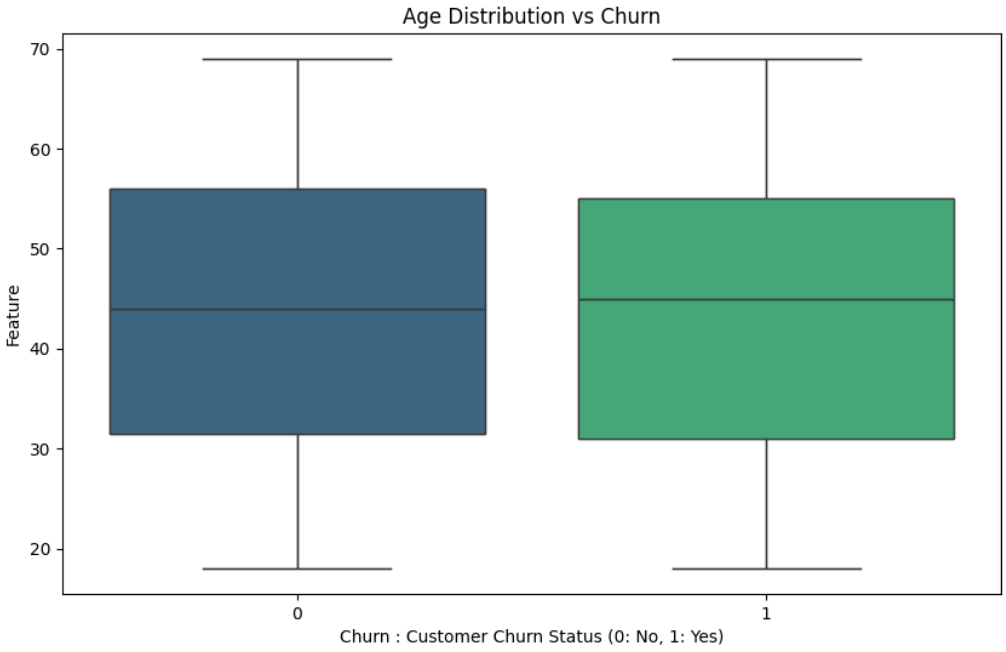
As the **number of calls increases**, the count of both churned and non-churned customers generally increases but with fluctuations. This could suggest that as customers need more support (and hence make more calls), the likelihood of them churning increases. However, the fluctuations indicate that this is not a strict rule and other factors might be at play.



From the graph, its clear that the count of both churned and non-churned customers varies with the number of years. This could suggest that the likelihood of customers churning changes as they spend more years with the company.

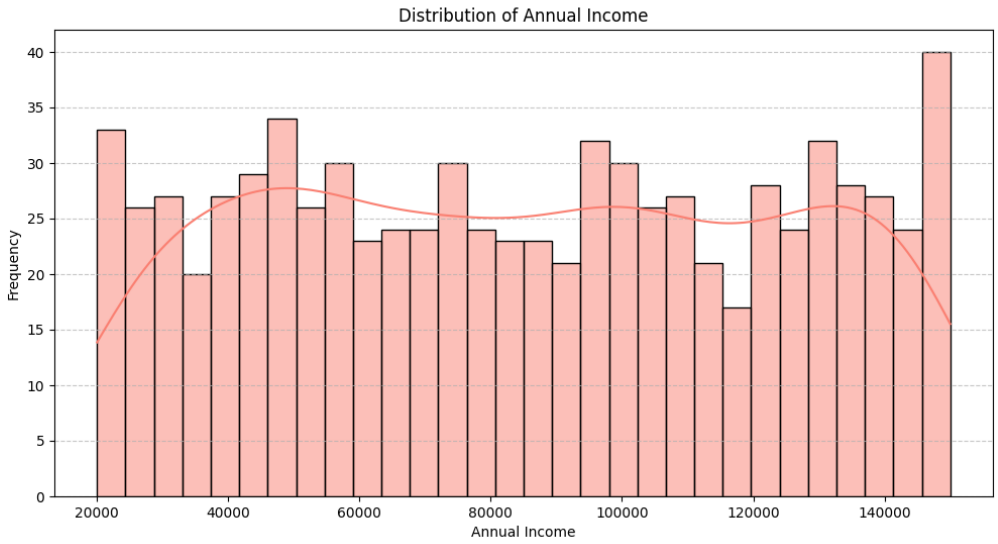
There are fluctuations in the count of churned and non-churned customers over the years. This indicates that customer churn is influenced by factors other than just the number of years a customer has been with the company.

There are three noticeable peaks in frequency around ages 25, 55, and slightly after 60. This could suggest that there are more individuals in the dataset who are around these ages. The trend line suggests that the frequency of individuals increases from age 20 to 55 and then decreases. This could indicate that the majority of individuals in the dataset are between these ages.

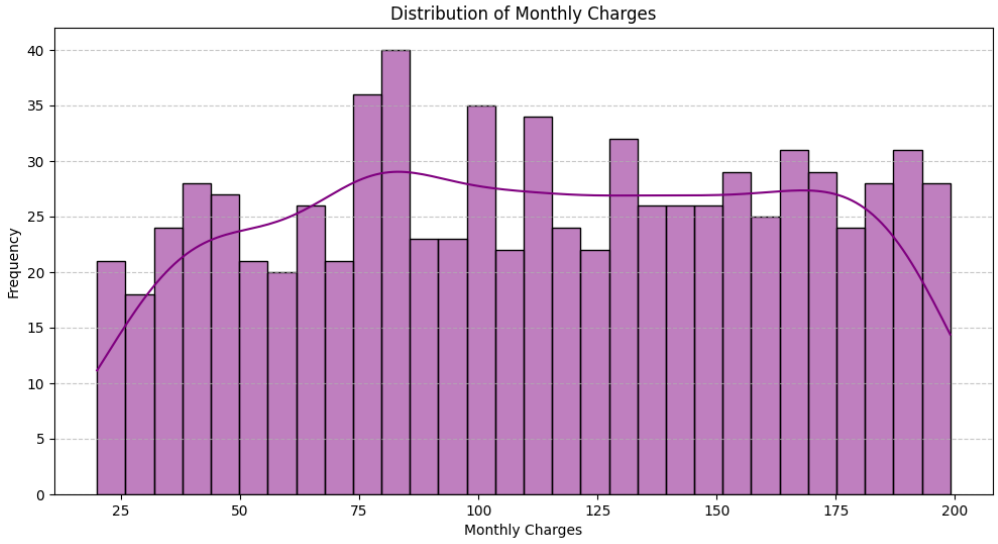


The **median age** is around 50 for both churned and non-churned customers, as indicated by the line within each box.
The **interquartile range** (the box) is similar for both groups, indicating that half of each group's ages fall within this range.

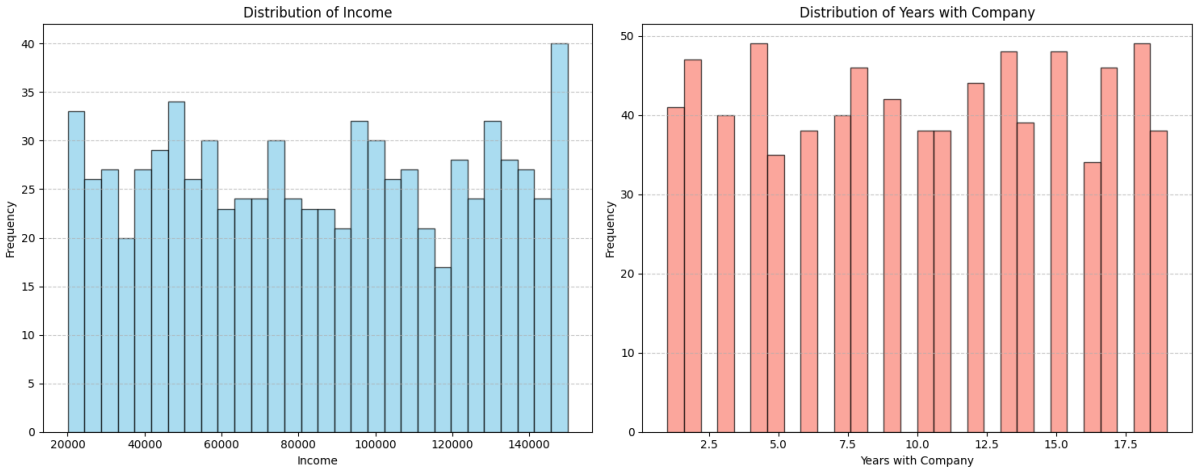
There are **no visible outliers** in either group, suggesting that the ages of customers in both groups are relatively evenly distributed without extreme values.



There are three noticeable peaks in frequency around annual incomes of 40,000, 80,000, and notably at 140,000. This could suggest that there are more individuals in the dataset who have these annual incomes. The distribution of income in this dataset is not uniform and is skewed towards certain income levels as per trendline.

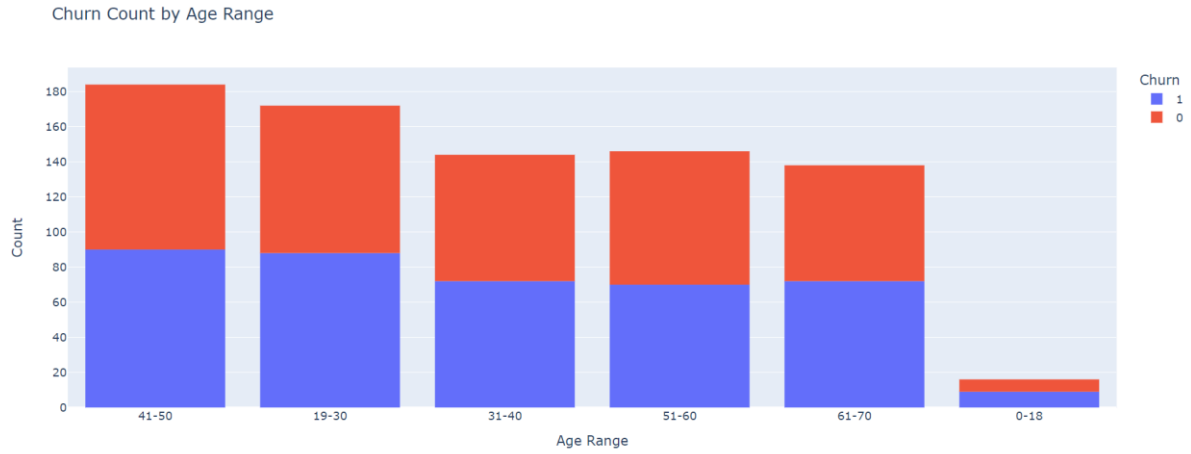


Frequency around monthly charges of 75 and 100. This could suggest that there are more customers in the dataset who are charged these amounts on a monthly basis. And distribution of monthly charges in this dataset is not uniform and is skewed towards certain charge levels.



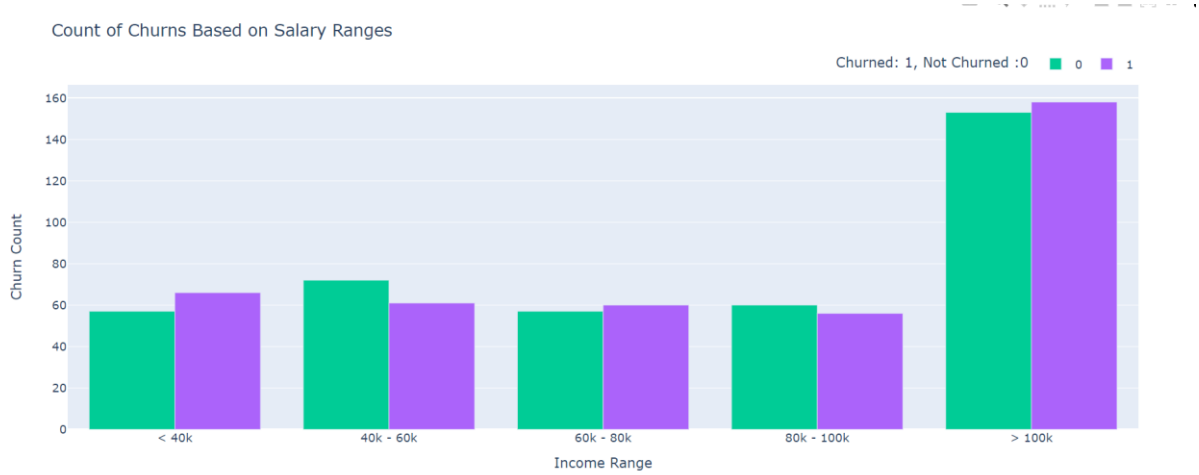
The income levels ranging from 20,000 to over 140,000 and bars indicate that incomes around 40,000 and just over 100,000 are most common.

Frequency of employees' tenure at a company ranging from less than 2.5 years to over 17.5 years. There is a notable increase in frequency at around 7.5 and again at around 12.5 years with the company.



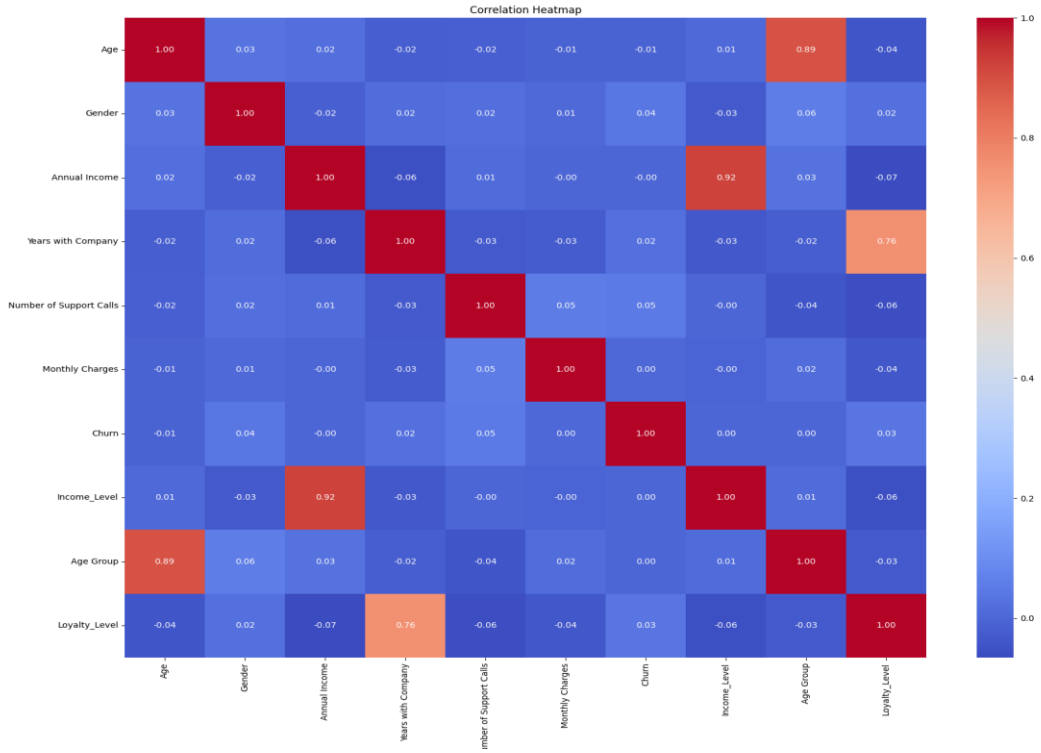
A few observations are as follows:

The age ranges **41-50, 19-30, and 31-40** have almost similar counts, with a slightly higher number of customers not churning if age range is **41-50**. The **51-60 and 61-70** age groups also show customer churn ratio is more or less same. There is a slightly high tendency to churn in the **0-18** age group. This could suggest that customers in this age range are highly likely not to continue doing business with the company. (Note : open image in a web browser for more interaction)

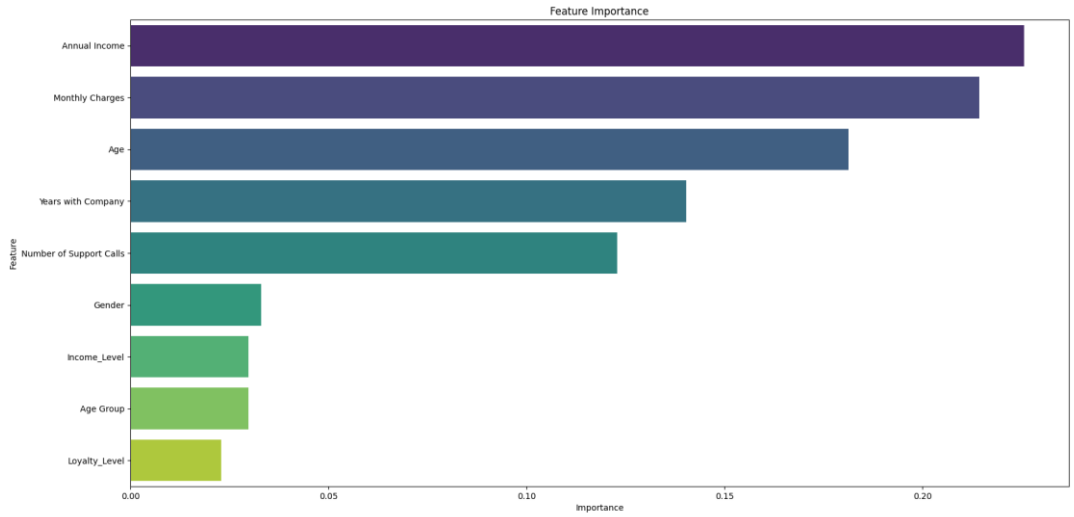


From the graph, we can observe that:

- Individuals with an **income range of >100K** have the highest churn count. This could suggest that customers with a higher income are more likely to churn.
- The **lowest churn count** is seen in the income range of **80K -100K**. This could indicate that customers in this salary range are less likely to churn.
- The churn count for the salary ranges **"40K -60K"** and **"60K -80K"** are almost equal. This could suggest that the likelihood of churn for these salary ranges is balanced.



It seems year with company and loyalty level are 76% correlated which is expected. There are no high correlated features among independent feature in dataset.



Annual income has major role in prediction of target variable.

Model Building:

Code: src\ml_modeling.py

Considering target variable as ‘churn’ and independent features as ['Monthly Charges', 'Annual Income', 'Age', 'Years with Company', 'Number of Support Calls', 'Gender', 'Income_Level', 'Age Group', 'Loyalty_Level'] for further machine learning modelling.

As an initial step trained KNeighborsClassifier machine learning model on train dataset for predicting the target variable (churn). Model has been evaluated on validation dataset.

Accuracy of KNeighborsClassifier is 0.9

Note : Script is scalable to add more ML model and compare the accuracy of different models. For now , consider only one model.

Hyperparameter tunning:

src\hyperparameter_tunning.py

Tunned selected machine learning model on validation dataset. Hyperparameters for each model has been configured as YMAL file.

Hyperparameter configuration for kNeighborsclassifier is as follow:

KNeighborsClassifier:
  model: "KNeighborsClassifier"
  param_grid:
    n_neighbors: [3,5,7,9,11,13,15,20,25]
    weights: ['uniform', 'distance']
    metric: ['euclidean', 'manhattan', 'minkowski']

Best hyperparameters after tuning KNeighborsClassifier :

[{"model":"KNeighborsClassifier","best_scores":0.94,"best_params":{"metric":"manhattan","n_neighbors":20,"weights":"distance"}}]

ML Model evaluation:

Code: src\model_evaluation.py

Use best hyperparameter of KNeighborsClassifier and test this model on test dataset

After apply Cross validation on trainset by using best tunned model :

Cross-validation scores:

CV Fold	**Value**
cv1	1.0
cv2	0.9875
cv3	1.0
cv4	1.0
cv5	1.0
Cv6	0.975
Cv7	1.0
Cv8	0.9875
Cv9	0.9875
Cv10	0.9875

Average cross-validation score: 0.9925 and Test set accuracy: 1.0

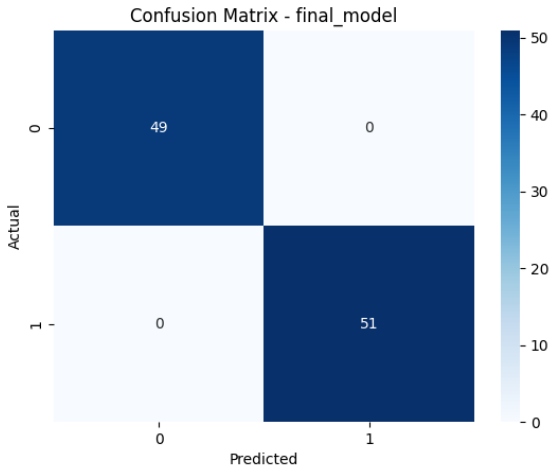
The cross-validation scores are very high and consistent, indicating that the model is performing consistently well across different subsets of the training data. An average score of 0.9925 means the model is accurately predicting the outcomes about 99.25% of the time on unseen data.

Final Model KNeighborsClassifier Evaluation Metric on Test dataset:

Metric: KNeighborsClassifier	**Value**
Accuracy	1.0
Precision	1.0
F1 score	1.0

These metrics on the test set are also perfect, with an accuracy, precision, recall, and F1 score of 1.0. This means the model correctly predicted all instances in the test set, making it an excellent fit for the data.

Confusion matrix for final model:



Summary:
The tuned KNeighborsClassifier model with the given parameters has shown exceptional performance:

It achieved an average cross-validation accuracy of 99.25%.
It achieved a perfect accuracy of 100% on the test set, indicating that it's likely a suitable model for the given data and can generalize well to unseen data.

Note: Cannot promise 100% accuracy in real world scenario

Based on the results of the predictive model for customer churn, here are some recommendations for the business:

Target High-Income Customers:

The churn rate is highest among customers with an income level greater than $100,000. The company should investigate the reasons behind this and implement strategies to retain these high-value customers.

Focus on New and Young Customers:

The 0-18 age group shows a higher tendency to churn. The company should focus on understanding the needs and preferences of younger customers to improve their retention rate.

Enhance Support Services:

The number of support calls is positively correlated with churn. This indicates that customers who require more support are more likely to churn. Enhancing the quality of customer support and addressing customer issues promptly can help in reducing churn.