



# Auto Insurance Churn Analysis

# Why Predicting Churn Matters in Insurance?

- Acquiring new customers is much **more expensive** than retaining the existing ones
- High churn leads to **lost revenue and higher marketing costs**
- Signals possible **issues with customer satisfaction or product offerings**



# Business Problem

- **How can we accurately predict which customers are likely to churn?**
- **What are the factors that lead to churn?**



# Impact of Understanding Churn Factors

- Enables proactive, personalized retention strategies
- Improves customer satisfaction
- Optimizes pricing models to reduce the churn



# Data Overview

- ~92,000 customers, 23 features
- Data types: numerical, categorical, date
- Key Feature Categories:
  - Demographics: DOB, age, education, marital status, children, home ownership
  - Financials: income, home market value, credit rating
  - Policy Details: annual premium, tenure, policy start date, acct suspended date
  - Geographic: city, county, latitude/longitude
- Churn Indicator: 1 = churned, 0 = retained



# Data Summaries & Data Quality Issues

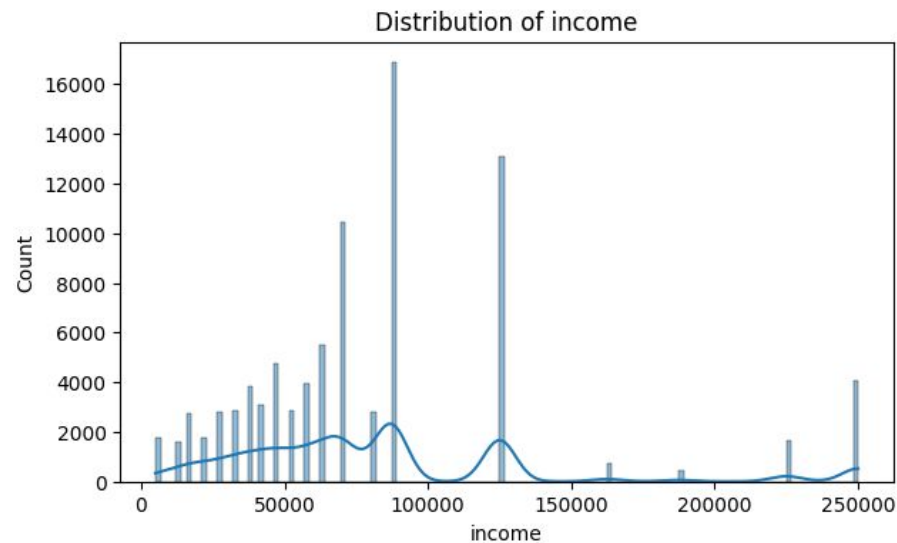
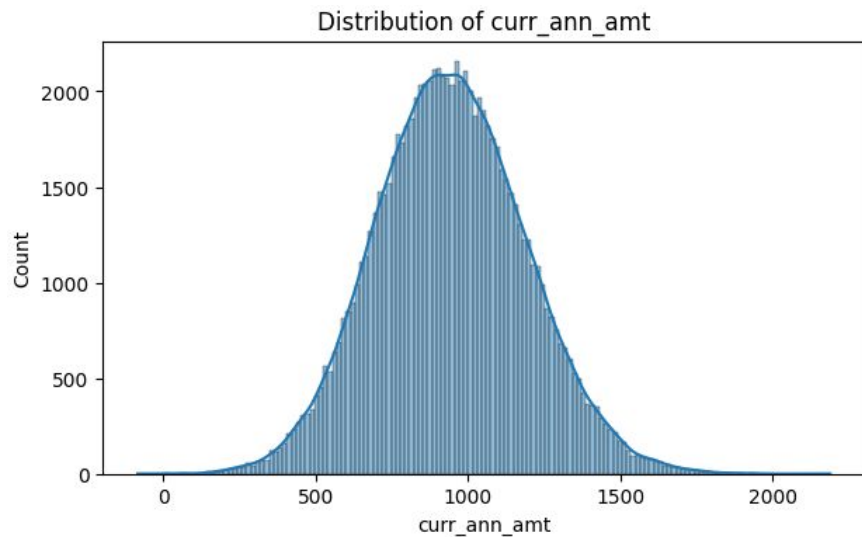
- Average tenure ~10 years  
(range: 20 days to 17 years)
- Average income ~\$81,500
- Average age ~56 years  
(max: 113 years)
- 84% have good credit rating
- Missing values in 'City', 'County',  
'home\_market\_value'
- Mean annual premium: \$940  
(wide range, min: -\$84)
- 82% homeowners
- 35% have college degrees
- 52% have children



# Handling Missing Values

VARIABLE	METHOD APPLIED
Longitude and Latitude (15%)	Fillna(city_avg_coord)
City and County (0.7%)	Fillna("Unknown")
Home_market_value (5%)	Dropna()
Acct_suspd_date (88%)	Drop the column
Address_id and Individual_ID	Drop the columns

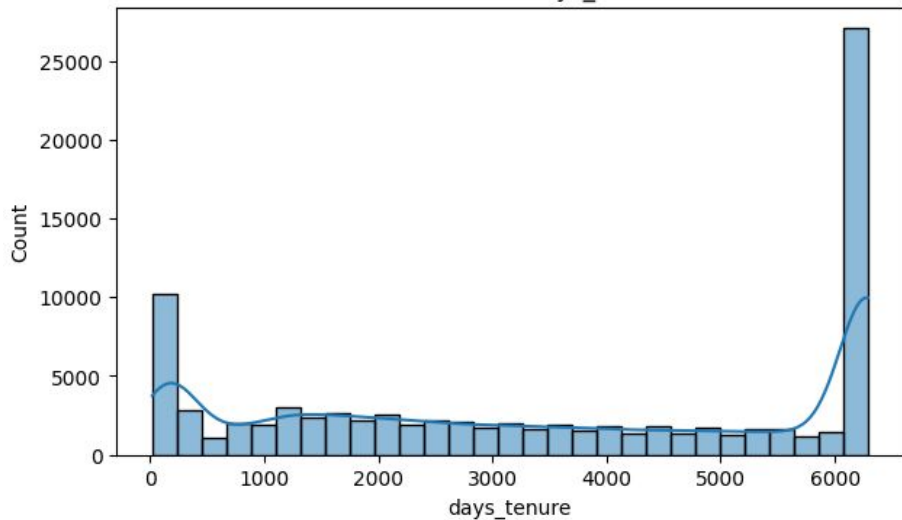
# Quantitative Variables Exploration



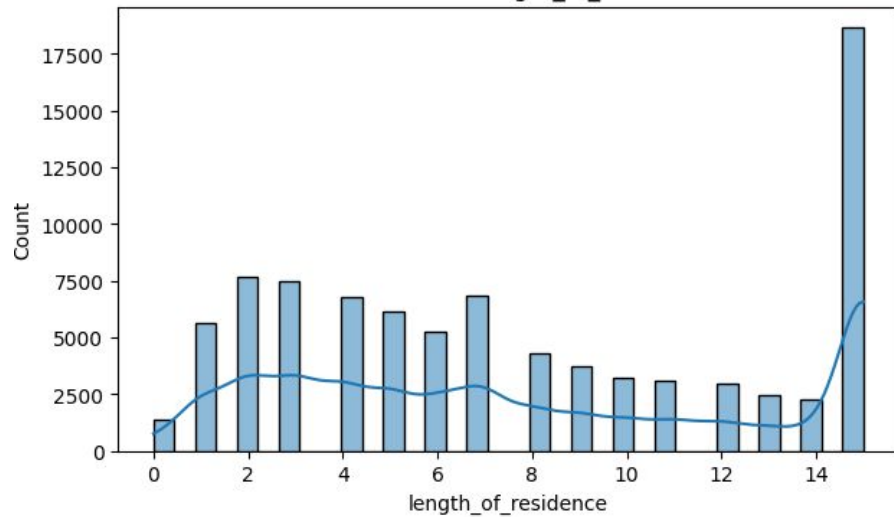


# Quantitative Variables Exploration

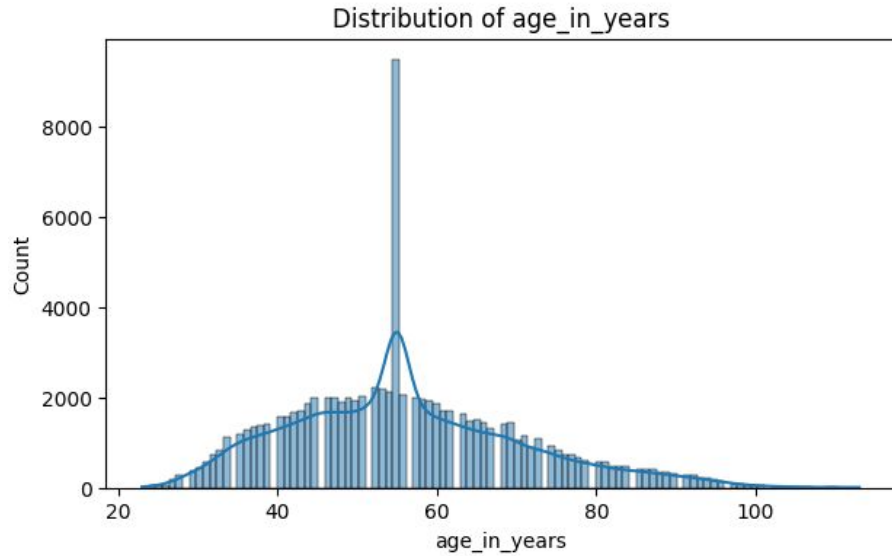
Distribution of days\_tenure



Distribution of length\_of\_residence

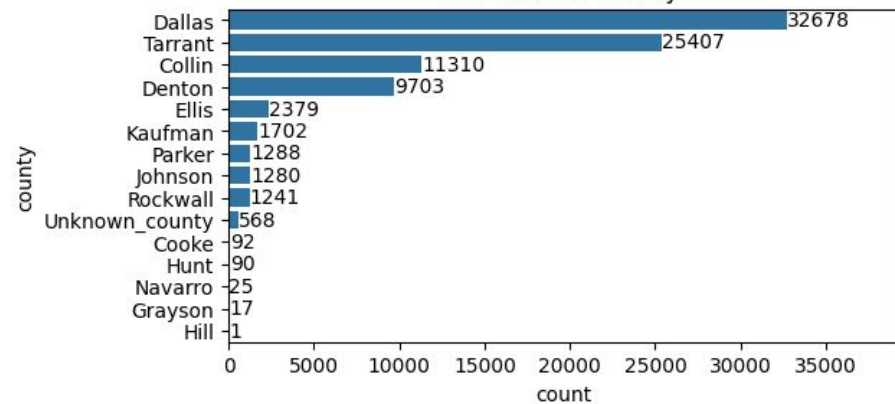


# Quantitative Variables Exploration

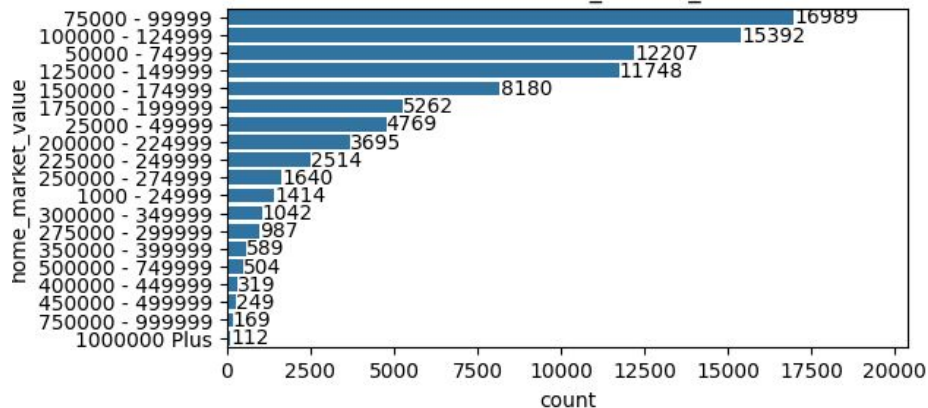


# Categorical Variables Exploration

Bar Chart of county

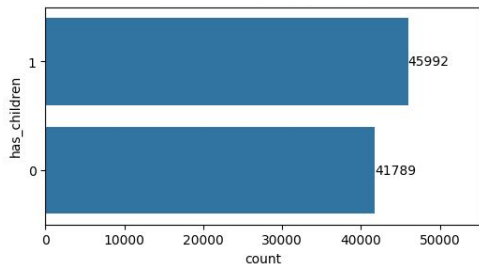


Bar Chart of home\_market\_value

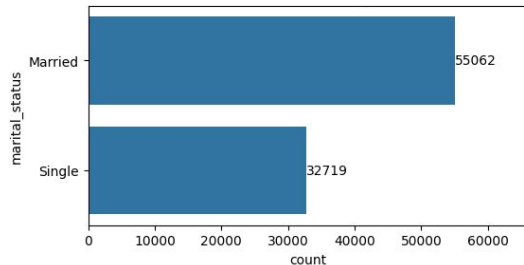


# Categorical Variables Exploration

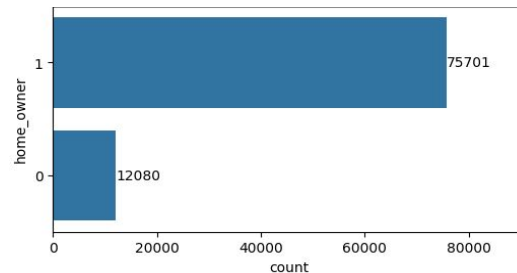
has\_children



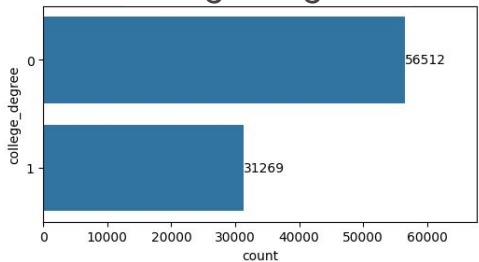
marital\_status



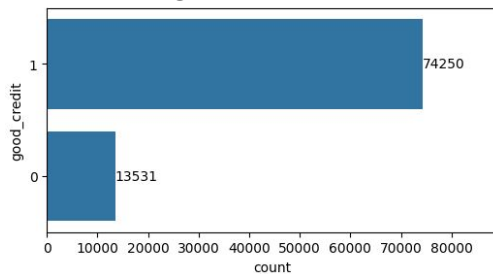
home\_owner



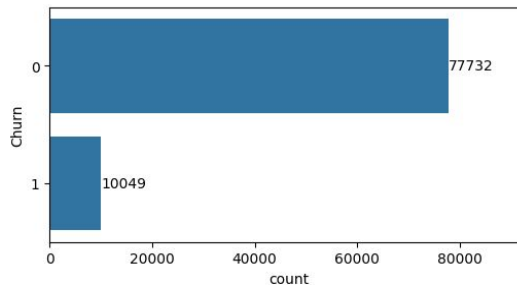
college\_degree



good\_credit

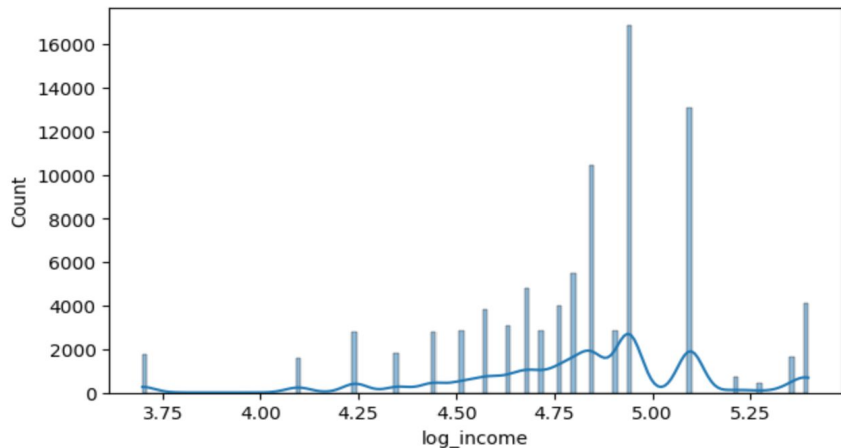


churn



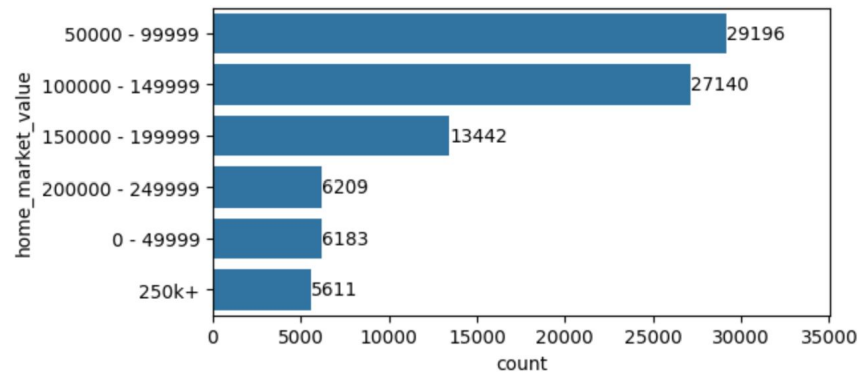
# Feature Engineering

Distribution of log\_income



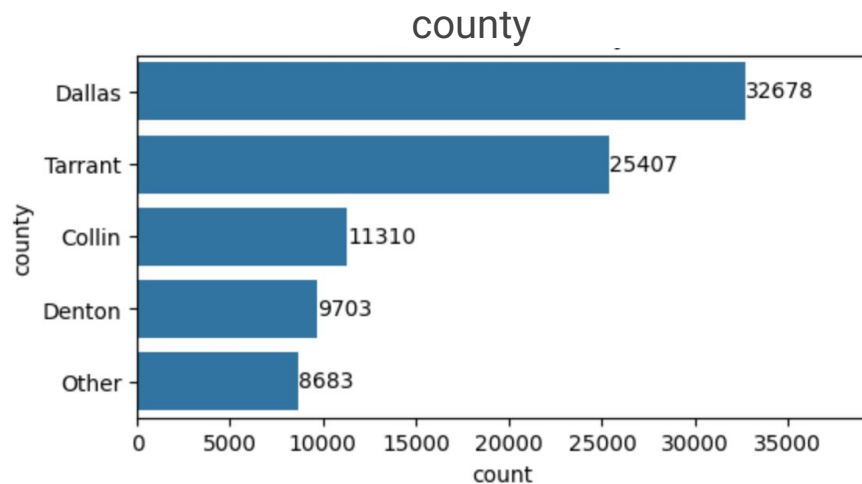
Applied **log transformation** to the income variable to reduce skewness and standardized the variable.

home\_market\_value

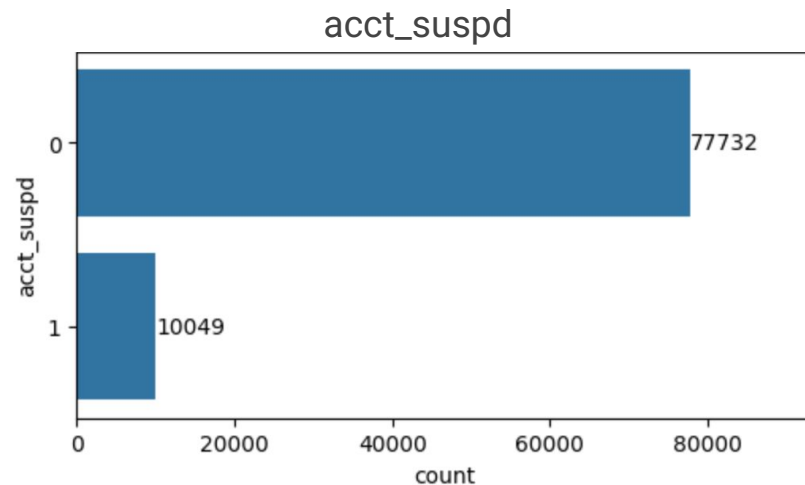


**Grouped(Binning)** home values into ranges to improve reduce the impact of noise or outliers.

# Feature Engineering

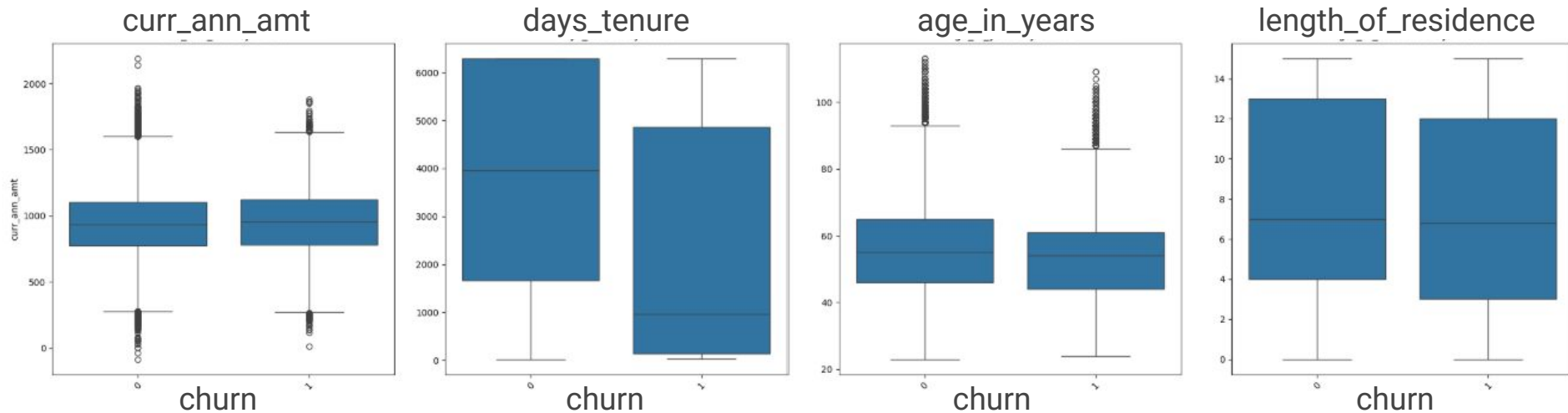


Retained top 4 counties; grouped others into a single "Other" category to simplify and better analysis



Created acct\_suspd column (1 = Yes, 0 = No) to enhance the readability.

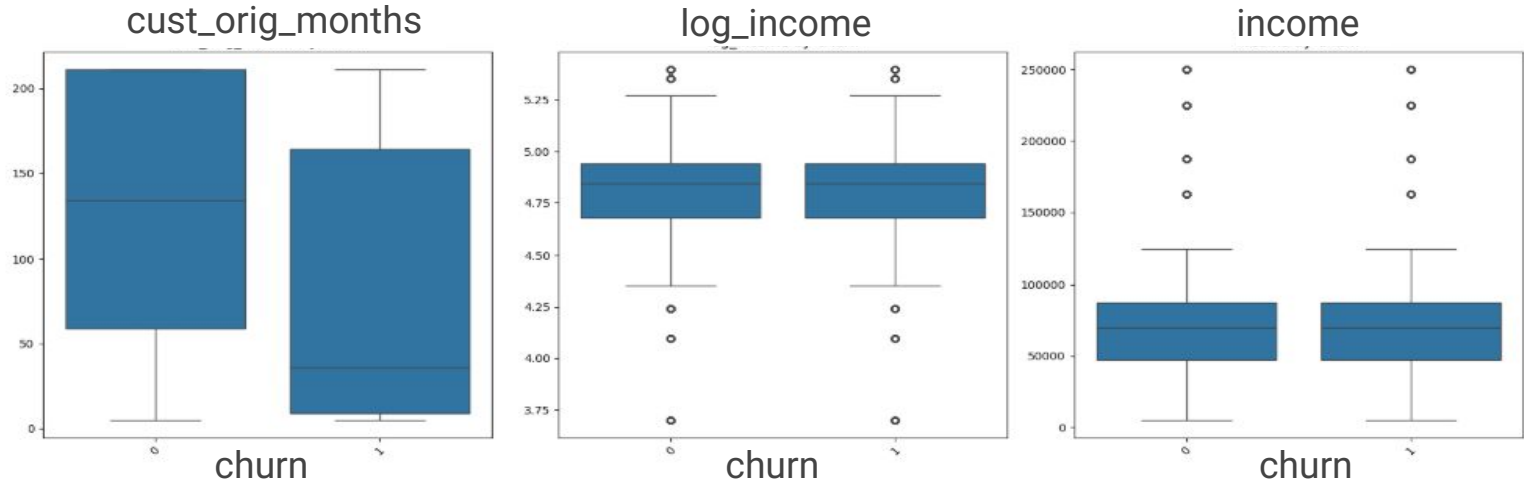
# Analyzing Predictors Using Boxplots



Variables like income, length of residence, age in years, and current annual amount show minimal impact on churn.

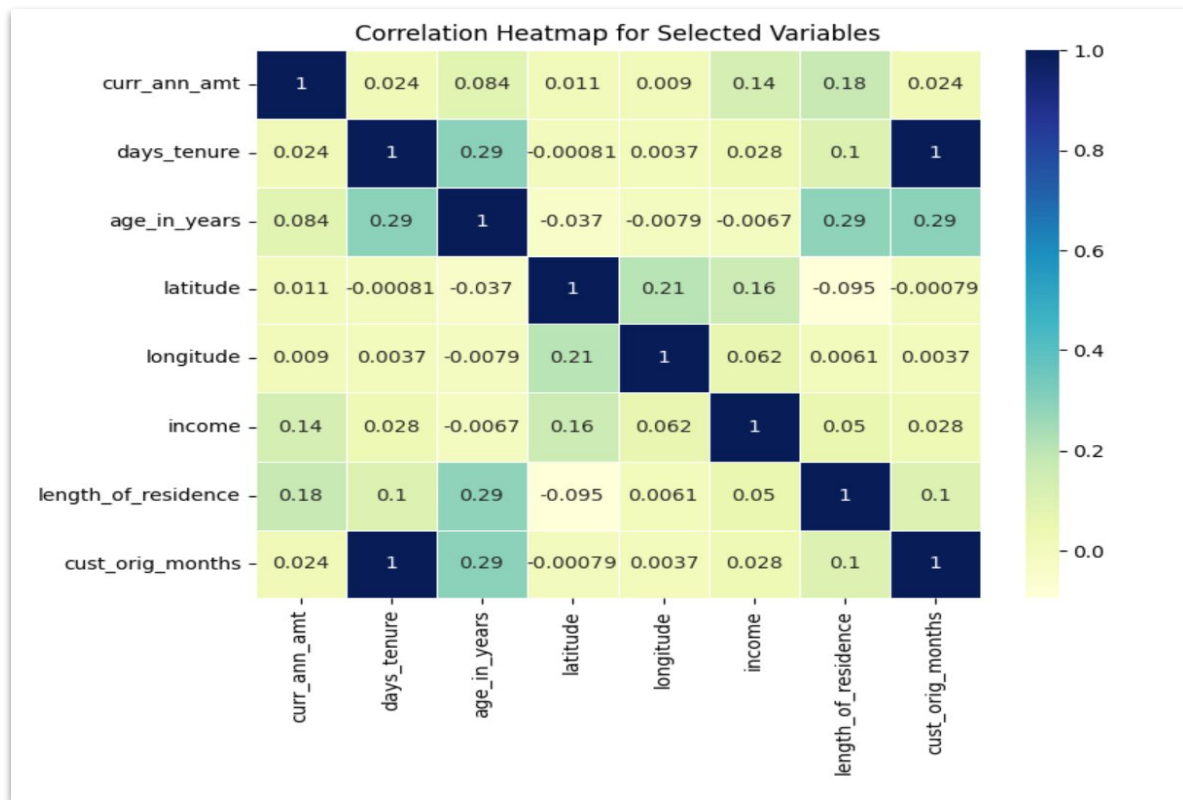
Tenure (cust\_orig\_months and days\_tenure) is a strong predictor of churn.

# Analyzing Predictors Using Boxplots





# Correlation Matrix



## Multicollinearity:

Strong correlation exists between **cust\_orig\_months** and **days\_tenure**.

**Action:** Dropped **cust\_orig\_months** to reduce redundancy and multicollinearity risks.

# Data Pre-processing

## Feature Selection

- Dropped redundant, non-informative, or highly correlated columns

## Data Transformation

- Created dummy variables for `home_market_value`, `marital_status`, and `county` to convert categorical data into numerical format for model compatibility.



# Data Pre-processing

## Standard Scaler

- Used `StandardScaler()` to standardize features by removing the mean and scaling to unit variance.

## Training and Testing data split

- 80:20 split, 80% of our data is Training Data and 20% of our data is Testing, test size was set to 0.2



# Classification Models Applied

- Decision Tree
  - SMOTE
  - Feature Importance
- Logistic Regression
  - Optimal Threshold
- K-Nearest Neighbors (KNN)
  - Hyperparameter Tuning
- Voting Classifier



# Decision Tree

Initially tested with various parameters:

- Depth: 6, 7, 8
- Minimum Leaf Samples: 500, 1000, 3000

Best results:

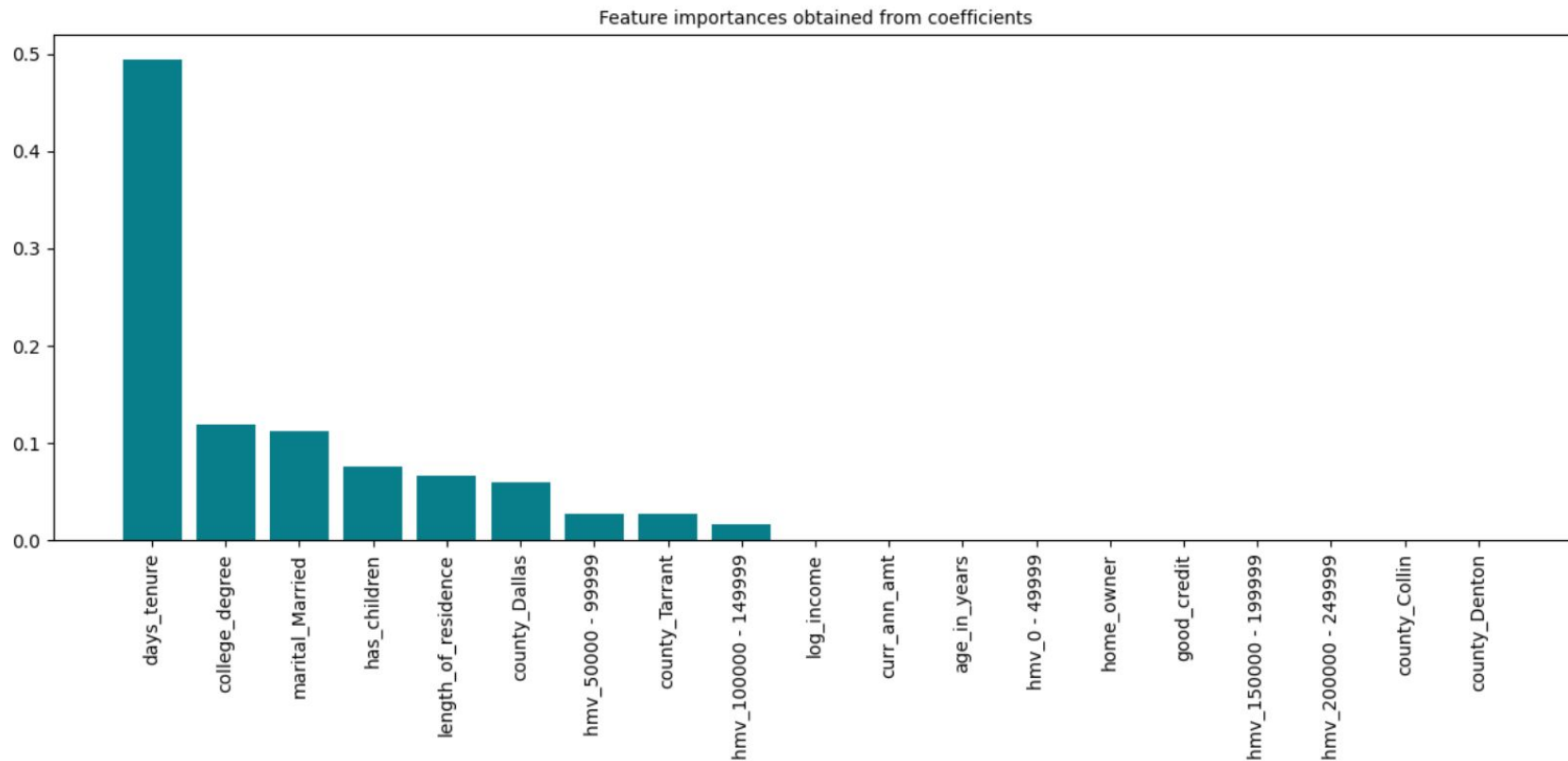
- Depth = 8, Minimum Leaf Samples = 3000

However, the tree suffers from *class imbalance* (1 = minority). So we applied SMOTE to address the issue.





# Feature Importances



# Logistic Regression

## Steps:

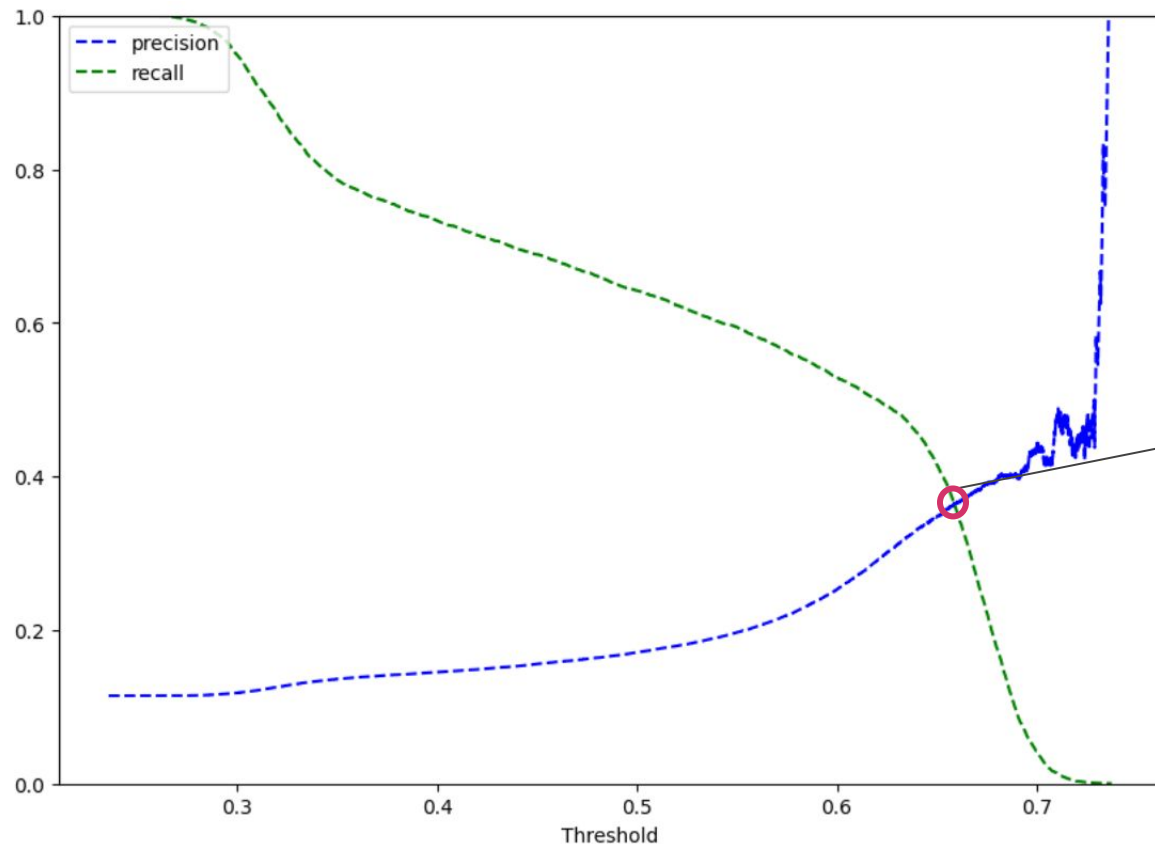
- Class\_weight parameter = “balanced” -> Addresses the class imbalance issue.
- 5-fold cross validation applied to assess model stability and performance variability.
- Achieved precision and recall of 0.16 and 0.66 respectively -> Large difference between the metrics.

Next, we find an optimal threshold to balance precision and recall.





# Precision vs Recall Curve for Optimal Threshold



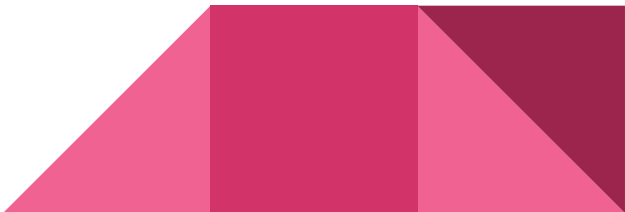
Threshold = 0.66

# K-Nearest Neighbors (KNN)

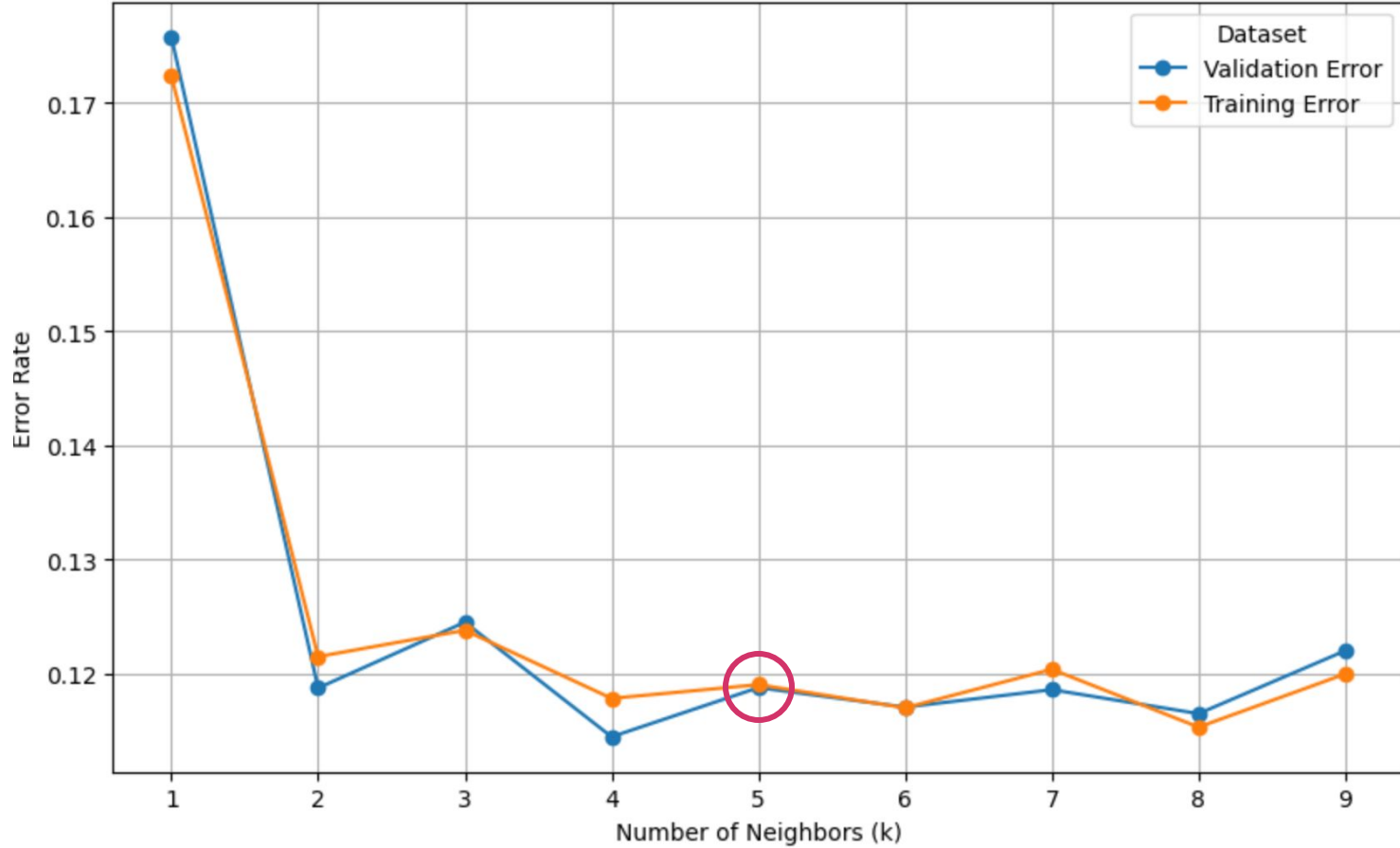
Steps:

- Loop over a few values of  $k$  to determine the best value.
- Plot the error rate vs number of neighbors ( $k$ ) graph for training and validation dataset.
- We chose  $k=5$  to best balance - error rate, overfitting and underfitting issues.

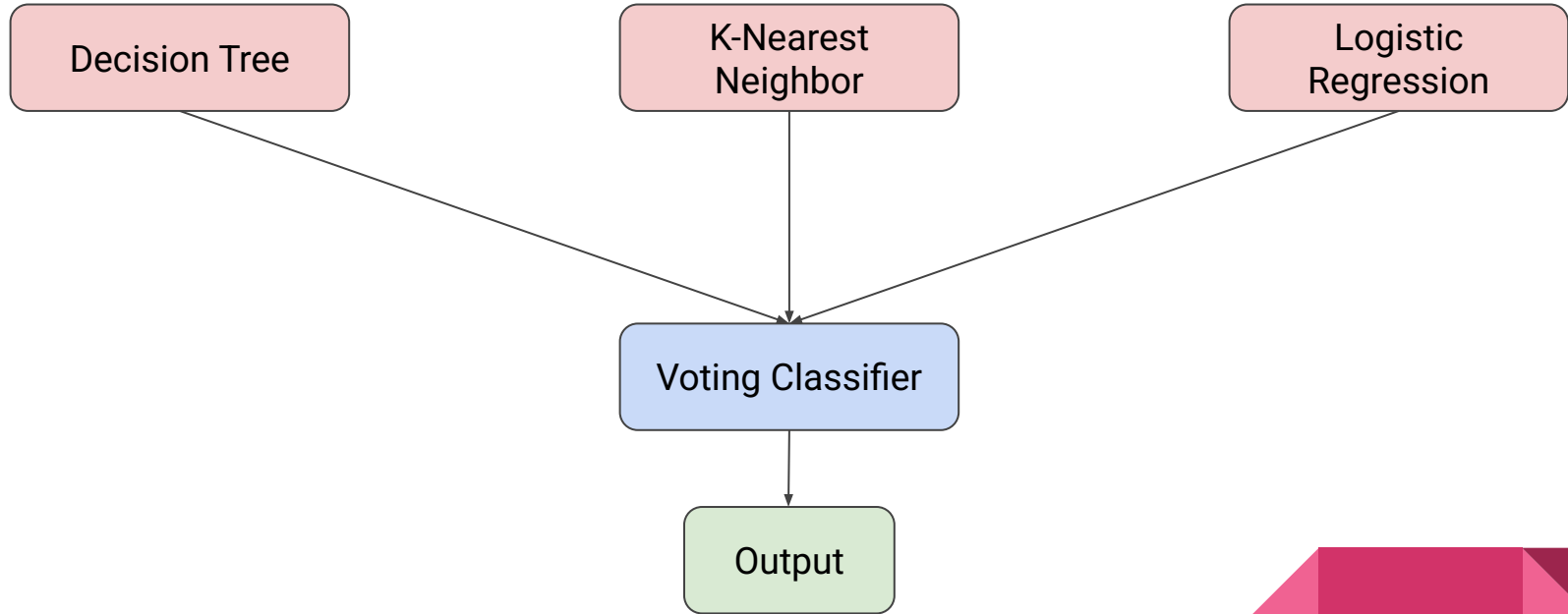
Next, we performed hypertuning to further improve the model's performance.



Error Rate vs. Number of Neighbors (k)



# Voting Classifier



# Comparing Model Performances (Class-1):

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.89	0.52	0.21	0.30
Decision Tree (With SMOTE)	0.82	0.32	0.47	0.38
Logistic Regression	0.60	0.17	0.66	0.28
Logistic Regression (Optimal)	0.86	0.37	0.36	0.37
K-Nearest Neighbors	0.88	0.44	0.19	0.26
Voting Classifier	0.89	0.50	0.33	0.39



# Comparing Model Performances (Class-1):

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.89	0.52	0.21	0.30
Decision Tree (With SMOTE)	0.82	0.32	0.47	0.38
Logistic Regression	0.60	0.17	0.66	0.28
Logistic Regression (Optimal)	0.86	0.37	0.36	0.37
K-Nearest Neighbors	0.88	0.44	0.19	0.26
Voting Classifier	0.89	0.50	0.33	0.39

# Comparing Model Performances (Class-1):

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.89	0.52	0.21	0.30
Decision Tree (With SMOTE)	0.82	0.32	0.47	0.38
Logistic Regression	0.60	0.17	0.66	0.28
Logistic Regression (Optimal)	0.86	0.37	0.36	0.37
K-Nearest Neighbors	0.88	0.44	0.19	0.26
Voting Classifier	0.89	0.50	0.33	0.39

# Conclusion:

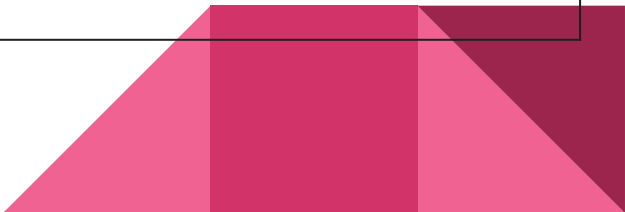
- Best model observed - Decision Tree with SMOTE
- Provides strong performance with a balance of precision and recall.
- Economically better and suitable for large datasets





# Business Recommendations:

Factors Affecting Churn	Impact on Churn	Strategy
Medium-to-high home values (\$50,000–\$150,000) and customers with children	Increases	Design personalized offers, family-centric loyalty programs, and retention campaigns
Longer tenure customers (over 5 years)	Decreases	Offer anniversary rewards, loyalty benefits, and exclusive retention incentives
Shorter tenure customers (under 1 year)	Increases	Provide introductory offers, proactive onboarding, and regular check-ins to build early loyalty
High proportion with good credit (84%) and college degrees (35%)	Decreases	Use tailored messaging and value-based offerings for creditworthy, educated individuals



# Business Recommendations:

Factors Affecting Churn	Impact on Churn	Strategy
Customers in lower-to-middle income range	Increases	Offer low-cost or value-based insurance products to appeal to price-sensitive segments
Recently relocated customers	Increases	Identify recent movers and engage them with personalized incentives and stay-benefits
Younger individuals and customers with shorter tenure	Increases	Implement onboarding programs, age-relevant engagement, and early financial incentives
High percentage of homeowners	Decreases	Promote home-related benefits, such as home insurance discounts or bundling offers



Thank you

