BANK CUSTOMER CHURN

A Reminder of the Importance of Context

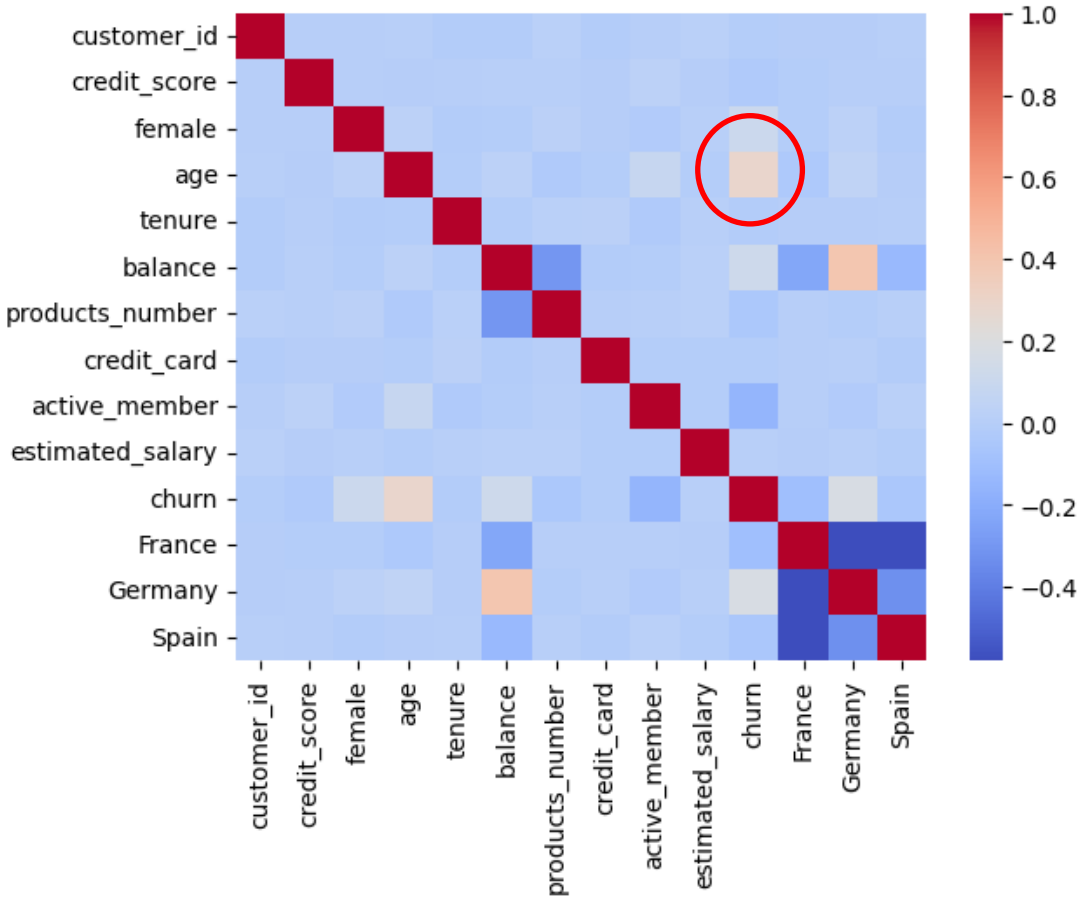
Churn Prediction

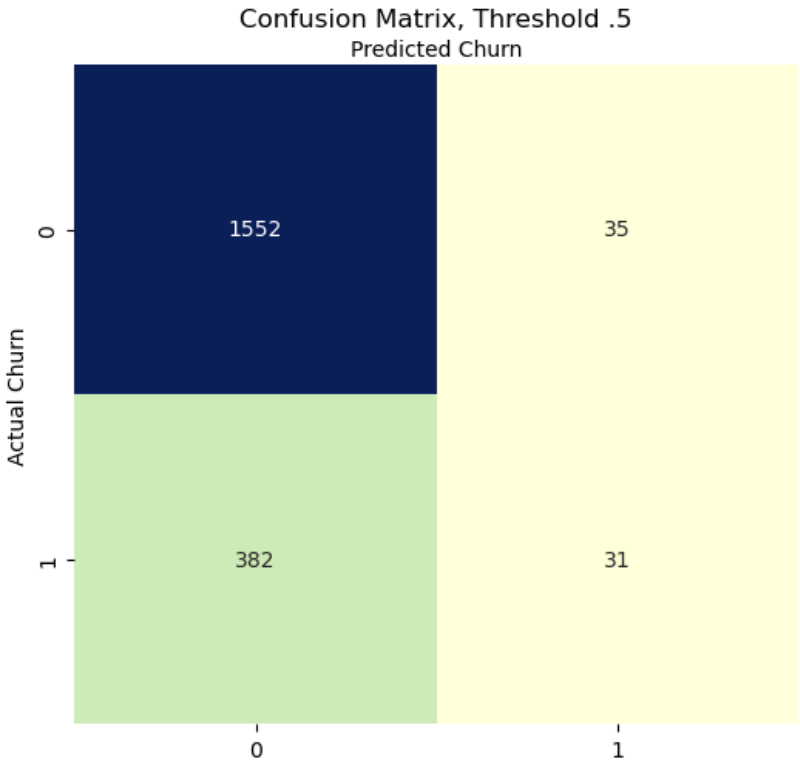
Churn is a measure of customer loss. This project began as an exercise in predicting customer churn through machine learning. ABC Bank wants to predict the which customers are likely to churn and uncover relevant insights from the data provided.

Data Source

Original data was obtained from Kaggle linked here. The data contains 10,000 rows and 12 columns. The target class, 'churn', is imbalanced with 2037 positive churn observations. No nulls or missing values were noted upon inspection, but should be sufficient for preliminary modeling. The dataset shows an overall churn of 2037/10000, 20.37%. This provides a baseline for predictions at ~80% accuracy assuming all predictions of "no churn". For any model to provide value, it must be more accurate than this 80% baseline.

Initial exploration of a correlation matrix provided the following results. With respect to churn, age was the only feature indicating moderate correlation (.285323). No other feature exceeded a ±.20 correlation coefficient with respect to churn. This is a relatively weak correlation value.





Logistic Regression Modeling

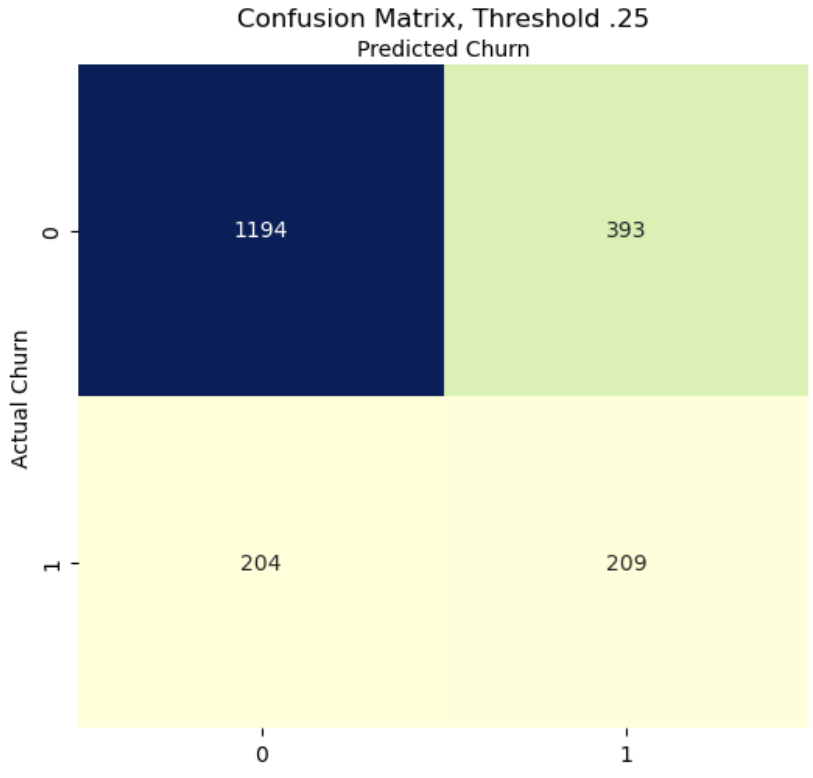
X and y data were defined. The variance between the balance and estimated salary variables compared with the single-digit variables necessitated scaling the data. Data was scaled using a standard scaler. Train/test split reserved 20% of the data for test. The SKLearn Logistic Regression model was applied to the data. Results were underwhelming with an initial mean accuracy score of .7915. A confusion matrix was constructed using the baseline threshold of .5.

A significant number of false negatives were noted in the initial modeling results. The business would rather include unlikely churners in marketing and strategic decision-making than miss any likely churners.

Accuracy v. Recall

Mean model accuracy is an important metric of any model. However, mean accuracy can lead to false confidence if other metrics are overlooked. As show (upper right), varying the decision threshold of predict_proba had little effect on the mean model accuracy. The same change in threshold had significant impact on precision and recall scores. Since the business wants to predict which customers are likely to churn, thresholding was tested to maximize recall (capture of true positives) while conserving as much model accuracy as possible.

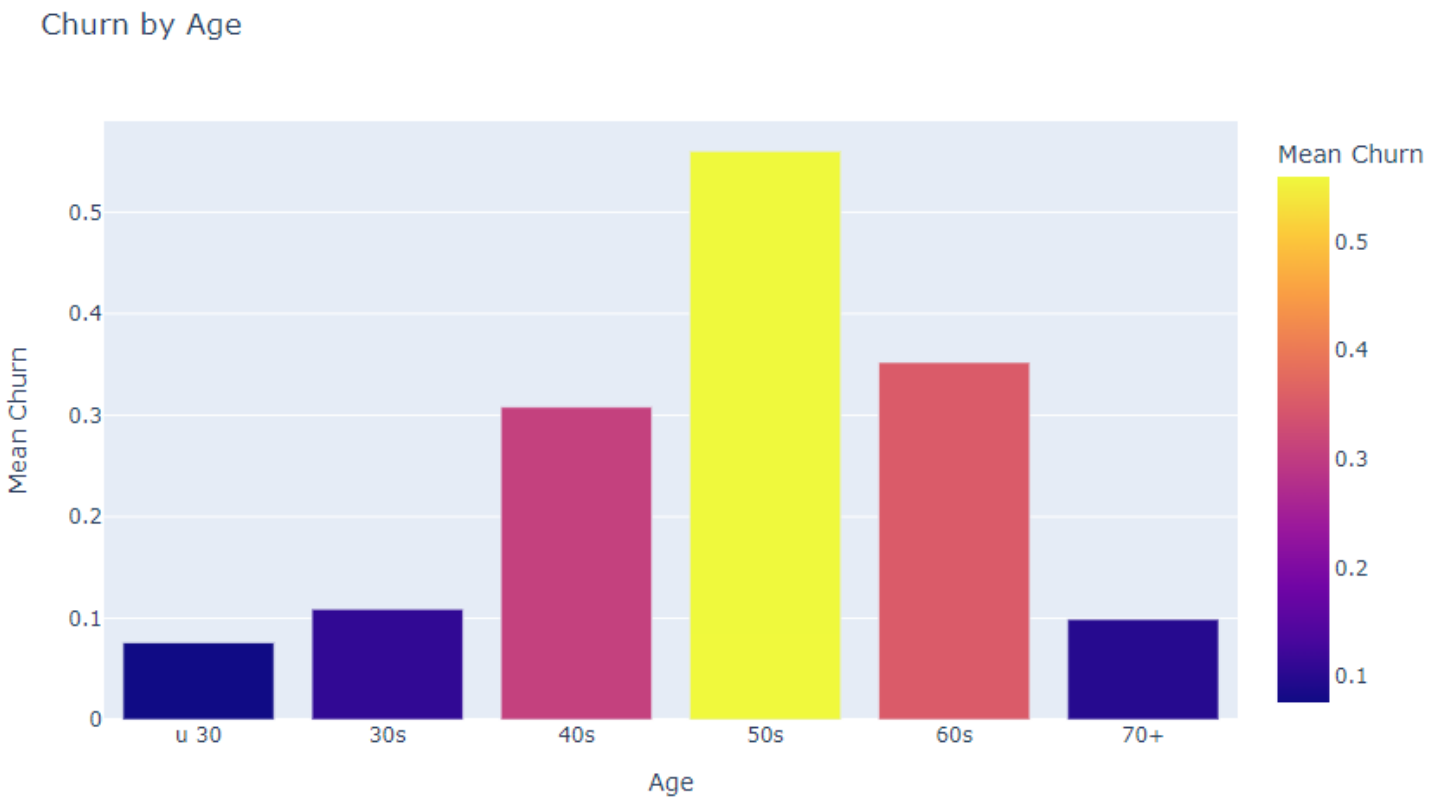
Losses in mean model accuracy are acceptable due to the importance of capturing as many likely churn customers as possible. False negatives are more significant than false positives for the purposes of this analysis. Therefore, recall was prioritized as a measure of accuracy. Decision thresholds were varied from .25 to .65 at an interval of .05 to maximize the recall score.



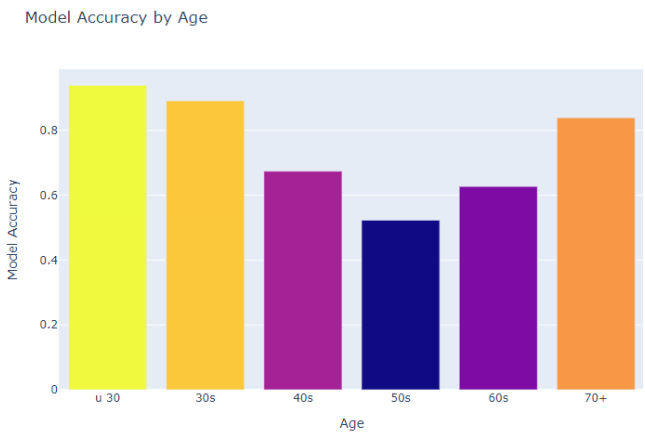
Age-Specific Churn

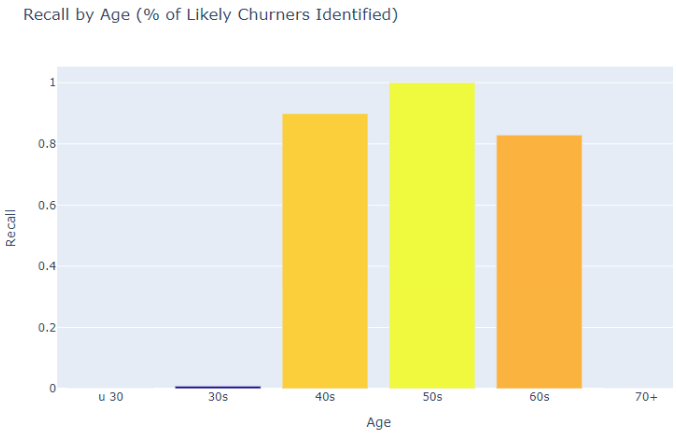
Early EDA revealed age was potentially correlated with churn. Age clusters were defined and subset from the original data. All customers under 30 were collected in a single age cluster (49 customers <20 years old). All customers aged 70 and above were also clustered (8 customers 80+). Other age-clusters were established in ten-year blocks (30-39, etc).

The largest cluster—30-year-olds— accounted for 3615 records, while customers over 70 accounted for just 99 records. Churn rates by age group were explored and visualized. Churn peaks in the 50-year-old cluster with a mean churn rate of .56 followed by 60-year-olds (.352) and 40-year-olds (.308). This establishes potential target groups for marketing and service offerings to reduce customer churn in these age clusters.



Models were run separately for each age cluster using the predetermined decision threshold of .25 to maximize recall. Overall model accuracy and recall scores were visualized by age cluster with surprising results. While mean model accuracy appears to drop with respect to the target age clusters, recall is significantly improved over the baseline illustrated previously. With the established decision threshold, the model is insufficient to make meaningful predictions. However, through the focused lens of targeted age clusters, the models effectively identify likely churn customers.





Insights & Caveats

The importance of context and the role EDA plays in providing context cannot be overstated. Context is the lens through which we view our problem statement and the framing device we apply to our data, methods, and metrics. Without a solid understanding of the requirements and the data, even the best models and conclusions can yield useless insights.

Modeling shows certain age clusters can be identified for churn with reasonable accuracy. This provides for reliability in targeting marketing efforts and new services to reduce customer churn within these age clusters.

With mean accuracy scores hovering near the mean, it is likely that other explanatory variables are missing from the currently available data. R^2 values are negative implying that the models are performing worse than the mean with respect to overall accuracy. This is to be expected following the reduced threshold required to increase recall scores. Within the limitations of targeted age clusters, the models are performing as intended. More features should be captured to improve model performance. The existing features are insufficient to explain the variance of the data.

Further Questions

- Cursory exploration of customer accounts with zero-dollar balances indicates a significantly higher churn rate than those with non-zero balances. Does this hold any explanatory power with respect to churn?
- Very few customers have credit scores below 500. If credit scores were bucketized, could similar target groups be identified?
- Are there services offered by competitors that are attracting our middle-aged customers away from ABC Bank? Are we collecting the reason for their departure from ABC Bank's services? Are we sufficiently educating our customers regarding our product and service offerings?