Fairway Bank Churn Reduction Project Report

Problem Statement

Fairway Bank is a multinational bank with locations in France, Germany, and Spain. In the next year, the bank will add branches in Belgium and Holland and plans to significantly update its online banking experience. In so doing, it must also renovate business practices to better align with customer needs—Fairway has seen a higher-than-average churn rate of about 20% this past year. Executives have turned to the bank's data science team to help achieve the goal of reducing churn to 10% by this time next year. To solve the puzzle of such consequential customer dissatisfaction, we will leverage customer data to build a predictive model that calculates churn risk per customer. We will also explore key features that are associated with attrition vs. retention. These insights will enable Fairway to identify customer profiles that tend to exceed a certain churn probability threshold. With the power of these data, we hope to make informed decisions on how to improve customer experience, subsequent satisfaction, and overall retention.

Data Wrangling

The data engineering team provided us with a CSV file containing a random sample of 10K Fairway customers and corresponding values across 15 features:

- credit score
- country (France, Spain, or Germany)
- gender (female or male)
- age
- tenure (# years that the customer has been with Fairway)
- balance
- number of products that the customer has purchased through Fairway (1 – 4)
- whether or not the customer has a credit card
- whether or not the customer is active
- estimated salary
- satisfaction score (1 – 5)
- card type (silver, gold, platinum, diamond)
- points earned by card use
- whether or not the customer has filed a complaint with Fairway
- whether or not the customer churned (target variable)

The data were clean with no missing values or duplicates.

Exploratory Data Analysis

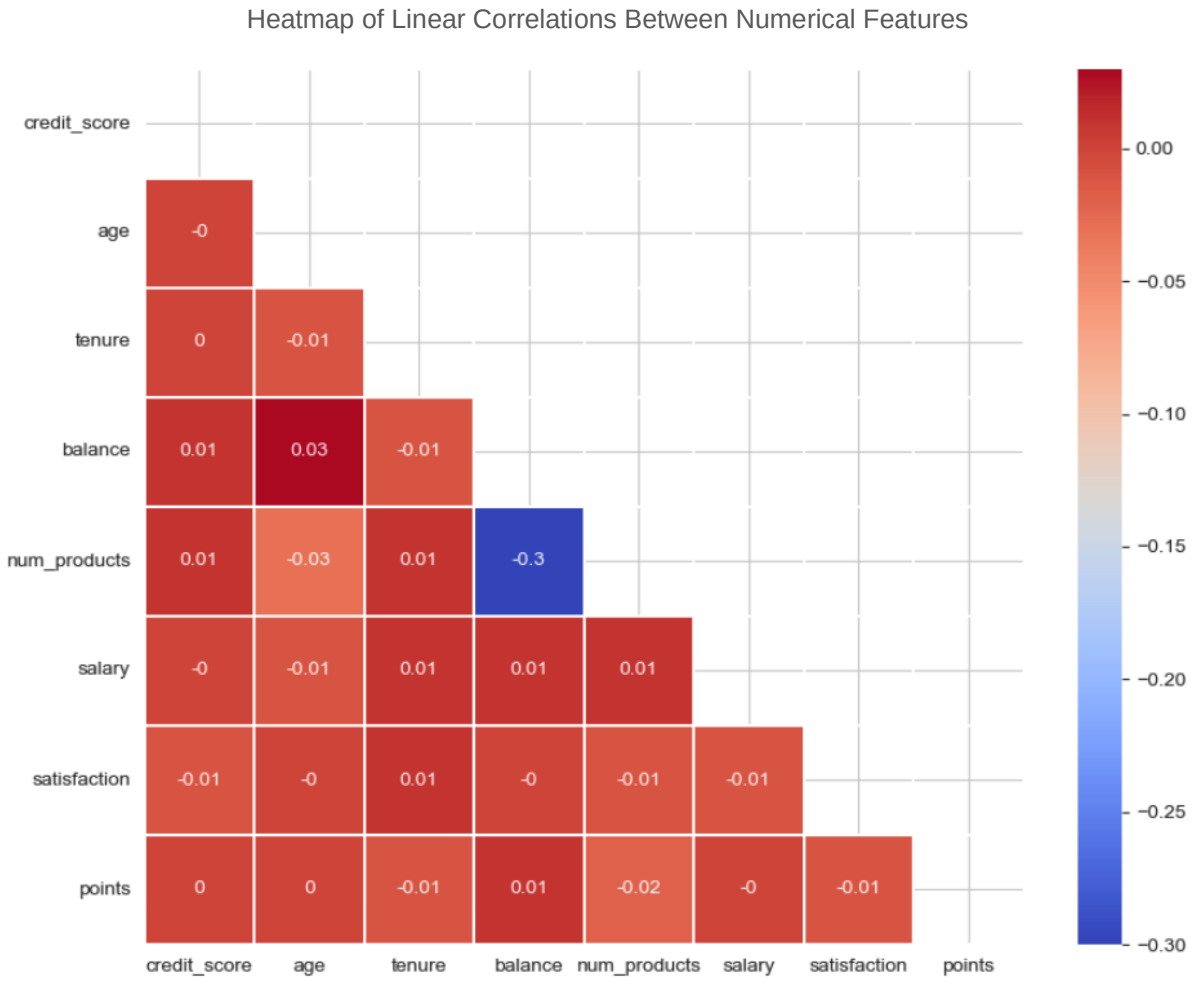
With the data clean and ready for exploration, we first looked at customer distributions across features.

Some initial takeaways:

- the data are imbalanced with a 4:1 retain to churn ratio
- the complain distribution looks identical to the churn distribution
- other noteworthy imbalanced classes include:
    - country (2:1 France to Spain or Germany ratio)
    - number of products (most customers have 1 – 2 products, few have 3 – 4)
    - credit card (3:1 has to doesn't have ratio)
- credit score and balance are bimodal with normal curves centered around 650 and an additional spike at the max of 850, and 150K with an additional spike at the min of 0, respectively
- the female to male ratio is about 45:55
- age is right skewed with a peak around 35
- tenure is uniform except for lower counts at the min and max (0 and 10 years)

- active, salary, satisfaction, and card type are uniform
- points is uniform except for a lower count at the min around 200

We next assessed linear correlations between numerical features.

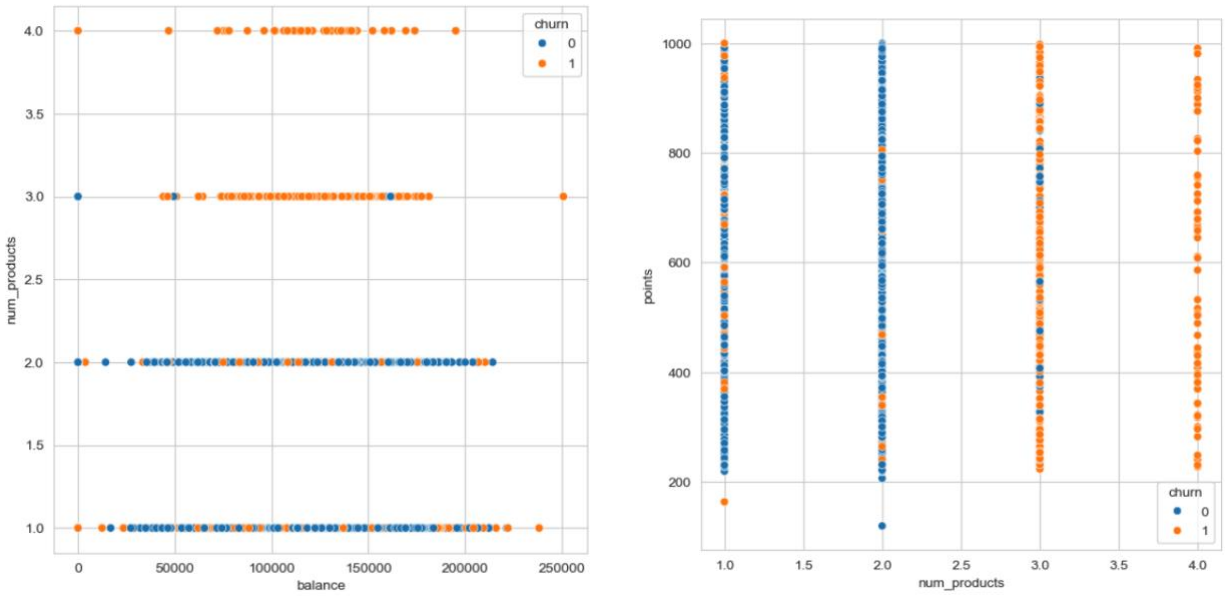


Number of products and balance have a weak negative linear correlation and all other features have no or very weak linear correlations. We proceeded to visualize pairwise relationships between features via scatterplots for numerical features and countplots for categorical features. We also added a hue for churn vs. retain to see if churned customers mapped to specific regions of the plot. While most plots were too noisy to interpret, a few demonstrated noteworthy indications. The clearest takeaways are illustrated in the following examples.

Scatterplots: Credit Score x Age, Age x Salary

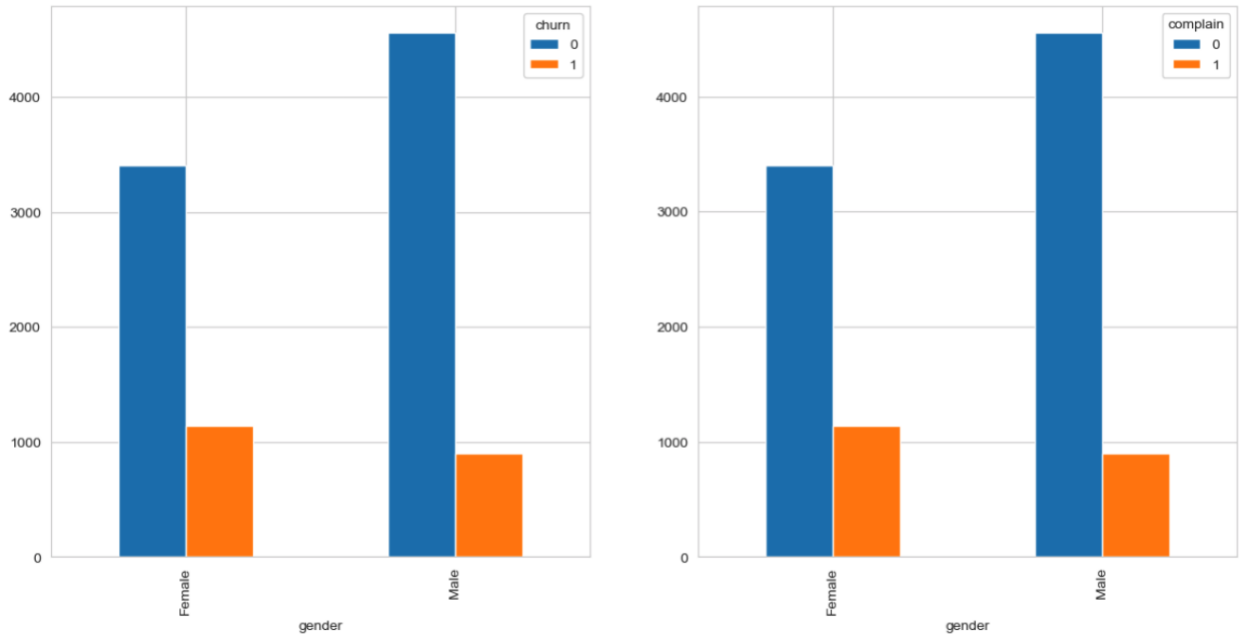
Churn is primarily clustered around customers between the ages of ~ 45-65.

Scatterplots: Balance x Number of Products, Number of Products x Points



Churn is primarily clustered around customers with 3 or 4 products.

Countplots: Gender x Churn, Gender x Complain



A higher proportion of females churned than males and more overall churns were female. The complain proportion looks the same.

Preprocessing and Cluster Analysis

We encoded all categorical values as 0's and 1's, with 1 column for binary features. For features with more than 2 categories, we applied dummy encoding to add additional columns per class (1 = is in the class that the column represents; 0 = is not in the class).

The next preprocessing step required our own judgement calls: we binned numerical features into a few discrete buckets. Bin sizes were based on business knowledge of relevant ranges for a given feature, and or relative to encompassing a large enough portion of the sample. In the case of our sample size of 10K, a few hundred observations per bin was sufficient, and several had a few thousand, depending on number of bins and the shape of that feature's distribution.

The bin ranges were:

- credit score[1]
  - 300 – 579 (Bad)
  - 580 – 669 (Fair)
  - 670 – 799 (Good)
  - 800 – 850 (Excellent)

- age[2]
  - 18 – 24
  - 25 – 34
  - 35 – 44
  - 45 – 54
  - 55 – 64
  - 65+

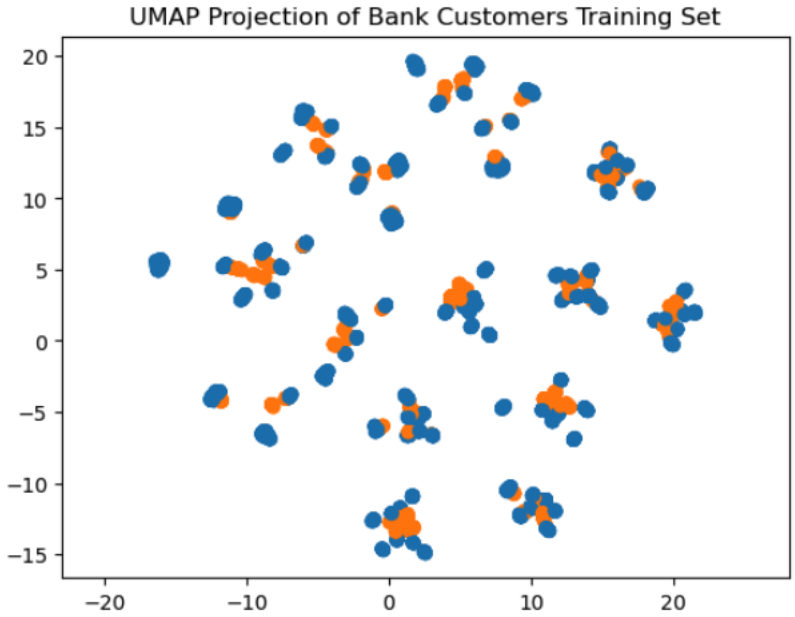
- balance
  - 0
  - 1 – (100K)
  - 100K – (150K)
  - 150K+

- salary
  - 0 – (50K)
  - 50k – (100K)
  - 100K – (150K)
  - 150K+

- points
  - 0 – 299
  - 300 – 399
  - 400 – 499
  - 500 – 599
  - 600 – 699
  - 700 – 799
  - 800 – 899
  - 900+

We proceeded to perform a cluster analysis and dimensionality reduction on the data. For these analyses we scaled the data using the Min Max Scaler to set values between 0 and 1. While a K-Means clustering seemingly delineated clusters based on country and card type, a Uniform Manifold Approximation and Projection mapped the features onto a 2-dimensional plane and demonstrated separability between churn and retain.

[1] https://time.com/personal-finance/article/different-credit-scoring-ranges/
[2] https://www.pickfu.com/demographic-segmentation150



We saved the UMAP embeddings for potential later use, however further application of dimensionality reduction on these data is yet to be applied.

Modeling Pt. 1

We applied a train/test split to the dataset, designating 80% of the observations as the train set and the remaining 20% as the holdout set. All model building and tuning would be done on the train set, reserving the test set only for final model evaluation, as this set is meant to emulate real world unseen data, e.g. existing customers that we have not trained the models on, and future customers. We considered the possibility of oversampling the minority class or undersampling the majority class so as to balance out the dataset, but given these techniques' cons of overfitting and information loss, respectively, we opted simply to use the present sample on hand and to stratify the target variable (churn), so that it was evenly represented in our train and test splits.

Next, we considered the following scoring metrics for our classification problem:

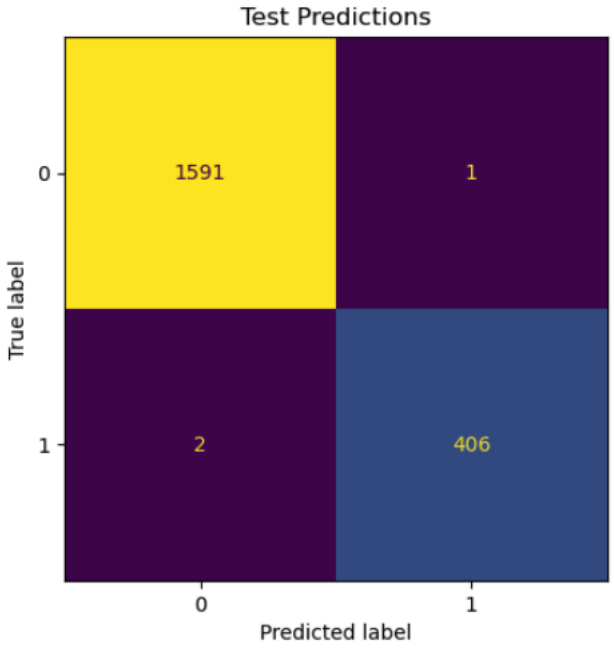
- accuracy: number of true predictions divided by total predictions
- precision: number of true positives divided by total predicted positives
- recall: number of true positives divided by total actual positives
- f1-score: harmonic mean of precision and recall
- ROC AUC: area under receiver operating characteristic curve, plotting the true positive rate against the false positive rate

Since our business goal is to correctly predict all churn cases, we want to minimize false negatives. In this case, the cost of a false positive is less than the cost of a false negative. We thus selected recall as our primary scoring metric.

We chose 10 classification models to apply to the train set:

- Random Forest
- Extra Trees
- Gradient Boosting
- Hist Gradient Boosting
- AdaBoost
- XGBoost
- LightGBM
- CatBoost
- Logistic Regression
- Support Vector

Before any model tuning, we compared the initial 'out-of-the-box' performance of each model. We applied a 5-fold stratified cross-validation to the train set and generated average recall scores across the folds. And lo, every model had a recall of 99.88%, except AdaBoost, with a recall of 99.63%. To confirm this almost-perfect performance, we plotted confusion matrices of model predictions on the test set. With the exception of AdaBoost, which had 3 false negatives, every model only had 2 false negatives and 1 false positive.



How could this be? We looked back to our exploratory data analysis and were reminded that churn appeared to be almost directly proportional to the complain feature. Moreover, we discovered that:

- only 4 / 2038 customers who churned didn't complain
- only 10 / 2044 customers who complained didn't churn

Whether or not a customer complained is a golden feature, and if a customer complains, Fairway has an immediate red flag for churn risk. This highlights the importance of improving customer support and complaint resolution. Before proceeding with further analysis, this should be noted as the top priority for Fairway. But can the bank be proactive and better support customers prior to any complaints, negating the customer's choice to complain and subsequently churn?

Modeling Pt. 2

We began modeling again with the same data, sans the complain feature. Upon calculating average recall scores for untuned models across the 5-fold stratified cross-validation on the train set, recall dropped by over 50%, with a median of 44.14%, XGBoost and LightGBM performing best with 47.42% recall, and Support Vector performing worst with 0.02% recall. It was time to make some adjustments.

The first approach we took was to adjust the probability threshold of each model. All models use the default threshold of 0.5. When a model predicts the probability of an outcome, if the probability is >= 0.5, the model outputs 1, indicating a positive prediction. Otherwise the model outputs 0, a negative prediction. By lowering the probability threshold, for example let's say to 0.3, we are adjusting model outputs to 1 for all cases where the predicted probability of a positive outcome is >= 0.3. Since a positive outcome here expresses a churn, we want to be conservative in our probability threshold so that even if there is moderate risk (about 30%) that a customer might churn, we label this instance as a churn prediction. The ideal probability threshold is another point to further discuss with stakeholders, but in the meantime, we explored ideal probability thresholds for each model.

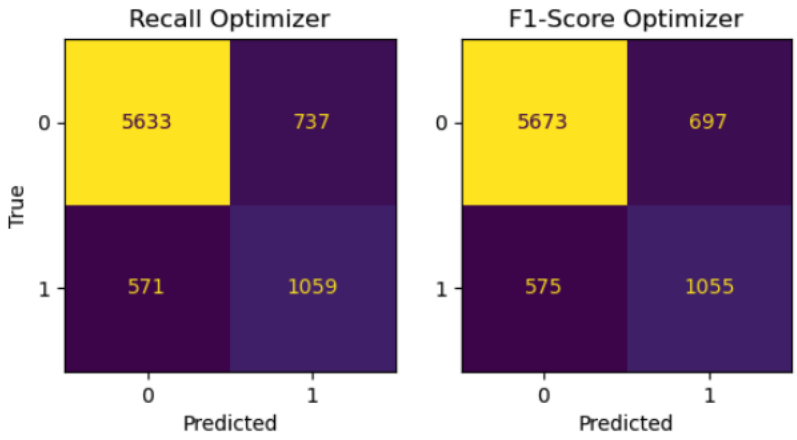
Identifying and assigning an optimal probability threshold for each model would not be as simple as tuning the threshold to maximize recall. This is because if we set a threshold to 0, all predictions would be classified as churn. The model would therefore have a perfect recall with no false negatives, because it would have predicted 0 negative outcomes. Thus, the false positives would be much higher and the model precision would suffer. To mitigate this imbalance, we used f1-score to tune probability thresholds, which returns the harmonic mean of precision and recall. This way we could find a probability threshold that minimized recall, without overcompensating with too many false positives. We ran the 5-fold stratified cross-validation on each model, with a grid search across probability thresholds (rounded to the hundreds place) ranging from 0.10 – 0.50, and determined the threshold per model that maximized the average f1-score across the folds. The optimal probability thresholds and their respective models are:

- 0.25 Logistic Regression
- 0.27 Support Vector
- 0.29 CatBoost
- 0.30 XGBoost
- 0.31 Hist Gradient Boosting
- 0.32 Extra Trees
- 0.33 Gradient Boosting
- 0.34 Random Forest, LightGBM
- 0.50 AdaBoost

Adjusting the probability thresholds accordingly improved model recall significantly, to a median of 60.50%, with CatBoost performing best with 63.31% recall, and AdaBoost performing worst (having maintained the original 0.50 probability threshold) with a 44.60% recall.

Next, we used Optuna to search for the best hyperparameters for each model. We optimized 2 versions of each model – 1 that selected hyperparameters that yielded the best recall, and one that yielded the best f1-score. We calculated confusion matrices for each of the of the respective model versions, across models. Predictions were accumulated from each train fold of a 5-fold stratified cross-validation on the train set, and compared to the true labels on the respective test folds. Overall, the model tuned to optimize f1-score was the favorable choice. An example of CatBoost's scores illustrates this.

Confusion Matrices for CatBoost Accumulative Predictions Across 5 Stratified Cross-Validation Folds



Though the recall is indeed slightly better on the recall-optimized version of the model, the f1-score-optimized version only has 4 more false negatives but also 40 fewer false positives. Even though false negatives are more consequential to our business scenario, in this case we are willing to trade a few false negatives for a noticeably larger gain in precision. Such was in fact the case across models, with AdaBoost even demonstrating better overall precision *and* recall on the f1-score-optimized version. We thus selected all models with hyperparameters tuned to maximize f1-score.

Modeling Pt. 3

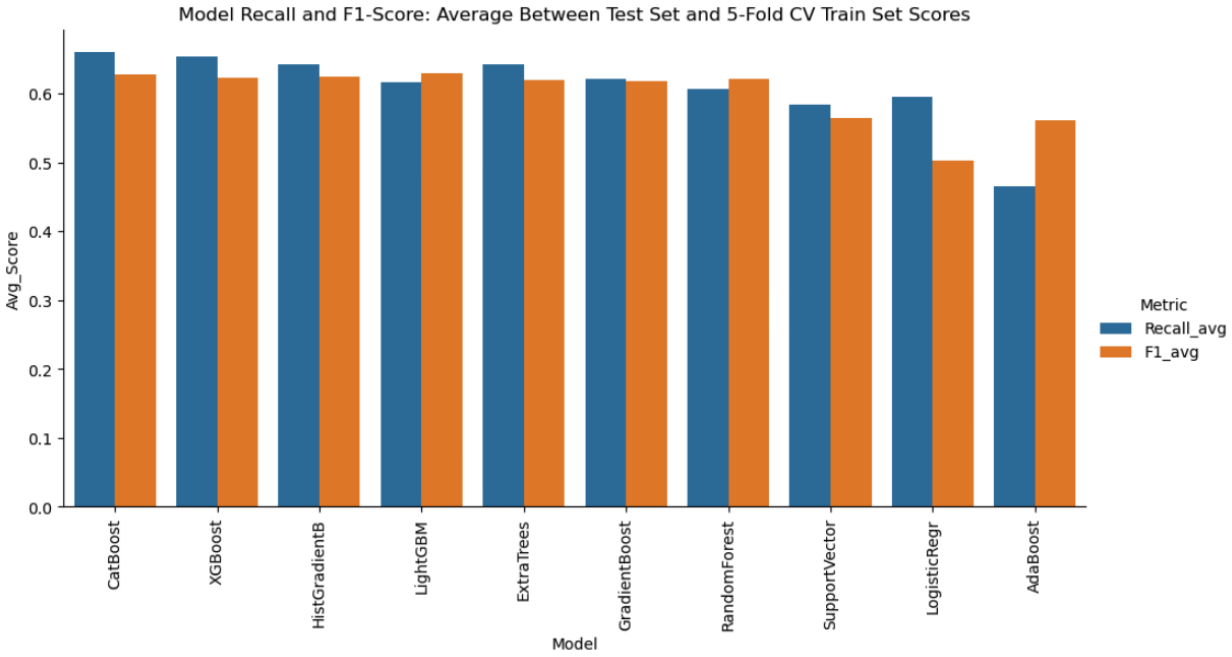
With all models tuned and appropriate probability thresholds determined per model, it was time to compare prediction results. For a robust overview of model performance, we charted performance metrics for recall, f1-score, precision, accuracy and ROC AUC. For each model, we performed a 5-fold stratified cross-validation on the train set and documented the mean score and the standard deviation of scores across the 5 folds for each metric. We also fit each model to the entire train set and calculated each of the respective scores on the test set. Upon comparing results across different metrics and data sets (average cross-validated train score or test score), we found that different models were top performers in different areas.

We decided to compute an aggregate score to reflect an all-around best model. To do this, we calculated a weighted sum of the respective metrics. First, we averaged all train and test scores per metric per model and subtracted each metric's train scores' standard deviation as a penalty for higher variation between predictions on different folds. We want the model to be as generalizable as possible. Next, metric weights were determined according to our business scenario. We prioritized recall in order to minimize false negatives. The second highest weight was assigned to f1-score, to ensure a healthy balance between false negative and false positive minimization. We then gave the third highest weight to ROC AUC, as this metric reflects the ability to classify between positives and negatives across varying thresholds. We assigned low weights to precision and accuracy, as the cost of false positives isn't as consequential to us, and the aggregate score's consideration for other true predictions over total predictions is taken care of in the preceding metrics. Respective weights and relevant metrics should be further discussed with stakeholders and calibrated accordingly. For this cycle of model evaluation, the following weights were used:

- 40% recall
- 30% f1-score
- 20% ROC AUC
- 5% precision
- 5% accuracy

We then selected our top 5 models for further evaluation. The models and their respective aggregate scores are:

- CatBoost 0.6798
- XGBoost 0.6795
- Hist Gradient Boosting 0.6726
- LightGBM 0.6650
- Extra Trees 0.6642

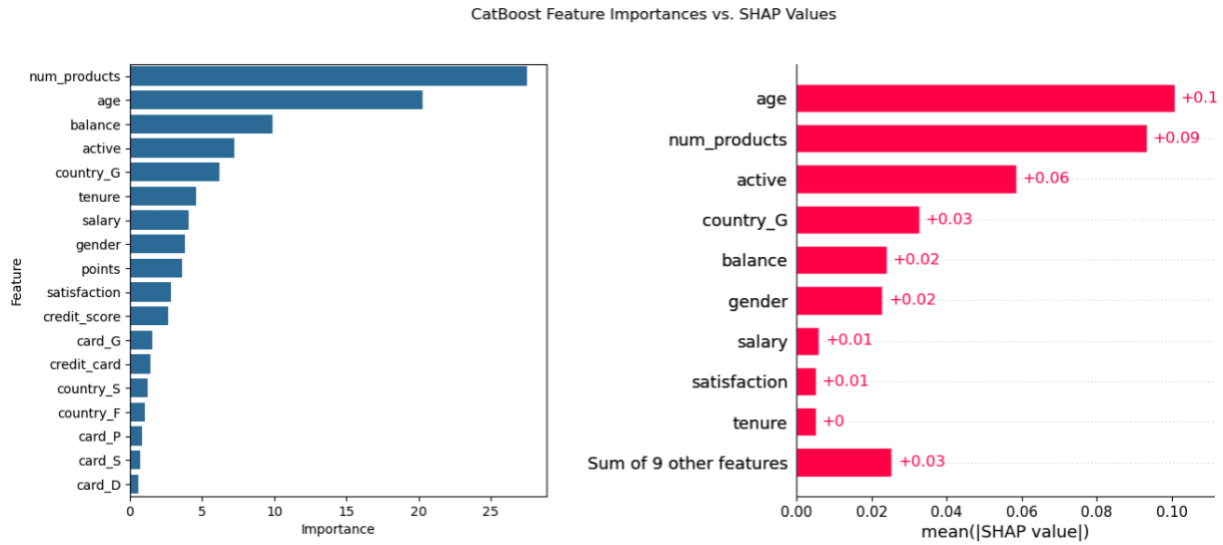


Importantly, the model with the top aggregate score has the highest recall, which is somewhat by design given that recall has the largest weight. We reiterate that overall, false negative reduction is our target, but it's important to consider other metrics to balance the strength of the model.

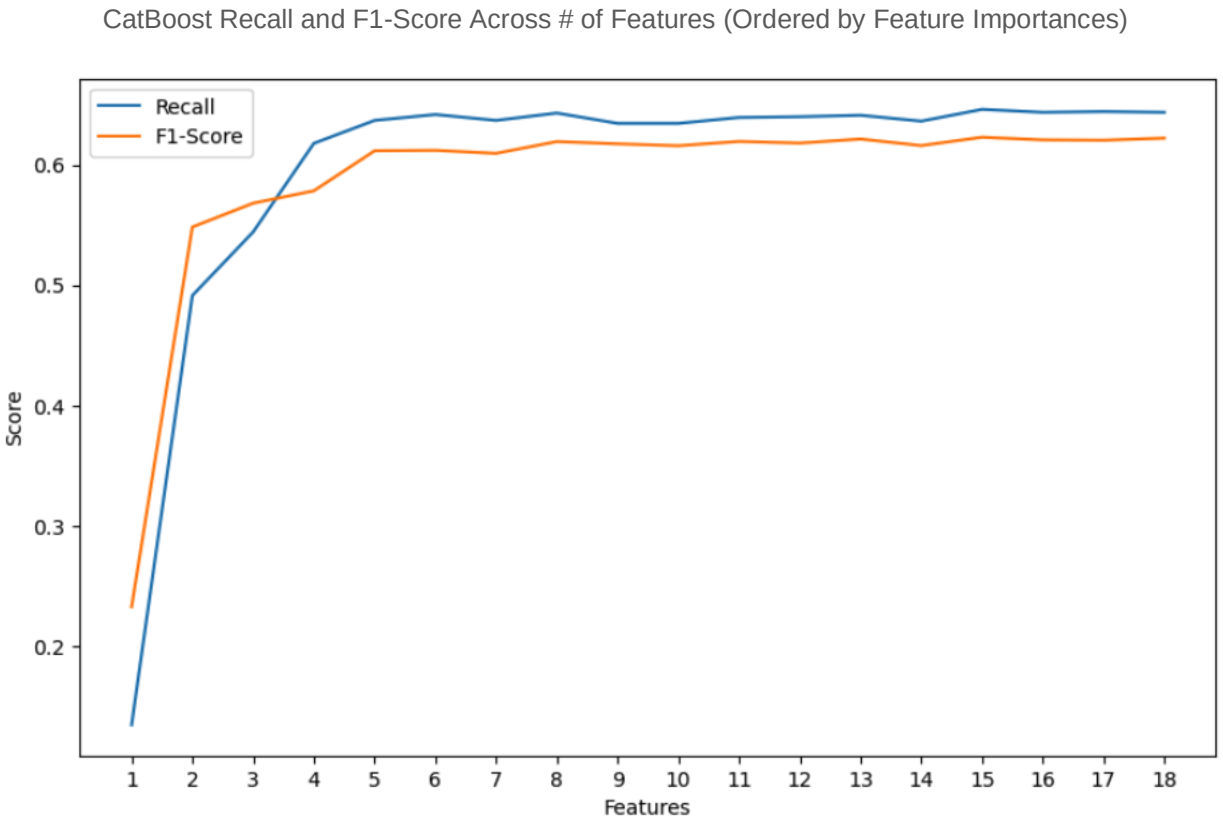
Another advantage to CatBoost is that it uses the lowest probability threshold of our top 5 models, classifying a prediction as a churn if it predicts a churn probability of 29% or higher. This ensures that the model is conservative in churn predictions. CatBoost was also the second fastest model, behind LightGBM, to fit and predict on 10K observations. Given its higher credentials in other areas, we still picked CatBoost as our top choice.

Finally, we assessed feature importances and Shapley (SHAP) values of our top 5 models. Feature importances are calculated based on each feature's impact on impurity reduction when it splits a node in a tree-based model. SHAP values compute how strongly each feature contributes to model output for a specific prediction and are calculated based on permutation shuffles of all other features. Overall, the feature importances and SHAP values of the top 5 models (Hist Gradient Boosting does not include a feature importances attribute) highlighted the following features to be of primary significance:

- age
- number of products
- active
- Germany
- balance
- gender



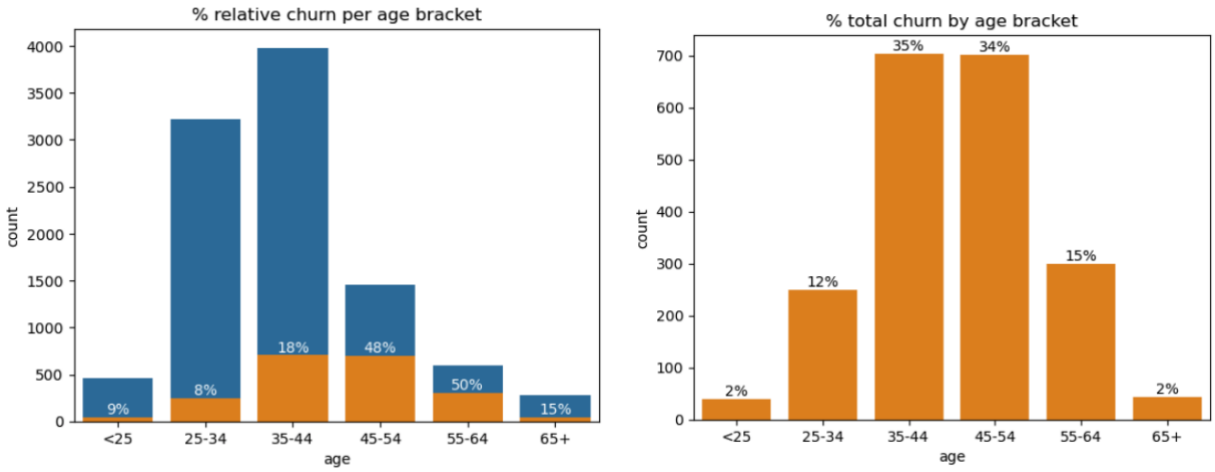
While the feature importances and SHAP values had some variation between one another and across models, they tended to highlight the aforementioned 6 features. To demonstrate how the features accumulatively contribute to model performance, the figure below indicates the recall and f1-scores for CatBoost when considering only 1 to all 18 of the features as ordered by feature importances.



There is a steep increase in performance with inclusion of the first 2 features, then a lesser but still steep incline to the 4th and 5th features, and after the 5th feature performance increase tapers off to a very gradual incline. In fact, performance peaks at 15 features with a cross-validated train recall of 64.60% and f1-score of 62.29%. These features excluding all card types except for Gold. Further analysis should be done and discussion had with stakeholders as to if any of the 18 features should be removed from the model. At the least, removing card type might be advantageous, both for model performance and for simplification of the input. Moreover, with only the first 5 features (number of products, age, balance, active, Germany), the model could be expected to generate predictions with an approximate 63.68% recall and 61.16% f1-score.

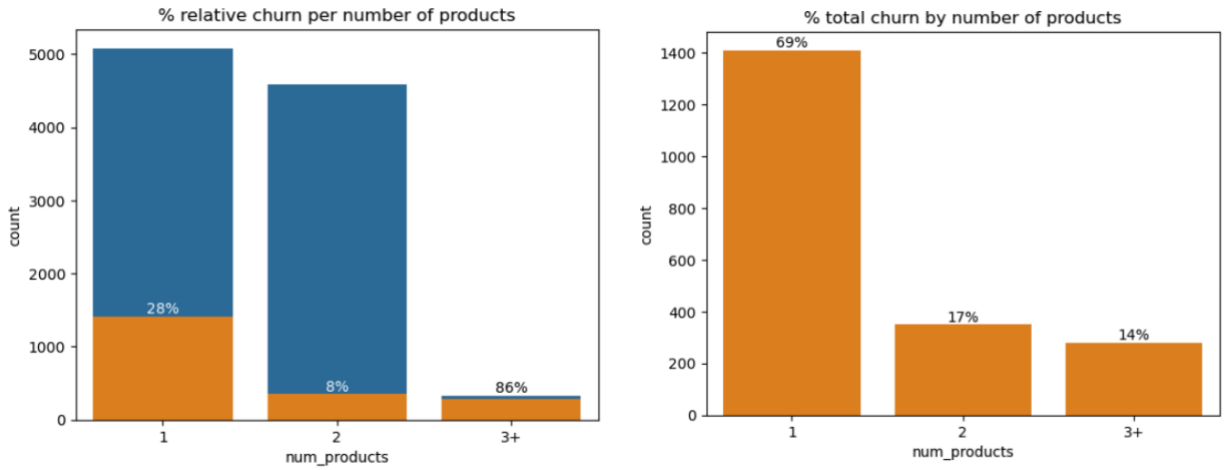
We also looked at how churn was reflected in the sample across each of our most important features. Below, we examine the top two: age and number of products. The figure on the left indicates churn as a proportion of the total cases in each category class. The figure on the right indicates total churns per class as a proportion of total churns in the dataset.

Countplots of Age Bracket and Churn Cases



The largest proportion of our customer base is 35 – 44. The majority of churns come from customers between 35 and 54. The 35 – 44 age bracket only has an 18% churn rate, but these make up for the majority of overall churn cases since this demographic represents more of our customers than any other bracket. On the flipside, while customers between 55 and 64 make up a smaller relative portion of our total customers, the fact that 50% of these customers churned translates to churned customers in this age bracket making up 15% of all cases. The only age range that currently fits within our ideal 10% threshold is customers 34 and under.

Countplots of Number of Products and Churn Cases



Most of our customers have 1 or 2 products but 14% of total churns come from customers with 3 products, as 86% of these customers churned. This indicates that Fairway should consider reducing product offerings to only 1 or 2 total products, and or further investigate why customers that go for more than 2 products seem to associate so strongly with churning. Customers with 2 products are our best-looking bracket, as only 8% of these churned, whereas 28% of customers with 1 product churned, accounting for 69% of total churn cases.

Future Implications

First and foremost, we recommend an assessment of Fairway's complaint resolution policy, as complaining is almost directly proportional to churning. If Fairway is able to develop solutions to mitigate customer dissatisfaction, many churns might be avoided. One suggestion is to allocate credit rewards as compensation if a customer reports an issue. Another possibility could be to integrate a direct call line for customers to discuss concerns with a representative. Person-to-person conflict resolution can be far more alleviating than going through an automated system. Fairway should also pursue more details on why customers are complaining, and seek to resolve the roots of problems.

Secondly, we press the need to evaluate churn likelihood, regardless of whether or not a customer complained. Herein we prescribe the CatBoost Classifier. The model suggests a probability threshold of 29%, indicating that any customer presumed at least this likely to churn is cause for concern. With the ability to identify 67.4% of churn cases (test set recall), the model could prove helpful in selecting customers that Fairway ought to strive to improve relations with. Inversely, the model could help select a target market for Fairway's expansions by identifying profiles of customers with high retention likelihoods. The model has been trained and tested using all aforementioned customer features, but a comparable performance can be achieved with as few as 5 features. Ideal performance metrics and potential feature inputs to include should be further discussed with stakeholders, with alternative model versions henceforth tested.

Finally, through our CatBoost model and the evaluation of other models with similar success rates, we identified several of the key features that seem to impact churn: age, number of products, whether or not the customer is active, whether or not the customer is based in Germany, balance, and gender. While exploratory data analysis of the sample indicated which classes of each of these features churned the most, additional hypothesis testing should be done to generalize these implications to the entire population.

Upon conferring with stakeholders, we hope to press forward with our predictive modeling and explore additional features, as well as additional classification models with a more thorough search of hyperparameter values. Other features of interest might include occupation, marital status, zip code, number of ATM uses per month, number of online banking logins per month, and number of bank visits per year. Additional models of interest might include K-Nearest Neighbor and Naïve Bayes. We also hope to revisit dimensionality reduction on the dataset sans the complain feature, and explore other possibilities for feature engineering that might enhance models' predictive power. In the meantime, we hope that the insights provided thus far can be useful and that the CatBoost Classifier grants Fairway a strong starting point in understanding customer churn likelihoods and developing actionable solutions to improve customer satisfaction and retention. Happy banking.