IBM Customer Churn

White Box and Black Box models

ZZSC5836 Data Mining & Machine Learning

Assessment 2: Project I

Alex Gould

1. Introduction

Customer retention is a key metric of our business; assuming our customers are profitable, more customers retained for longer translates directly into greater profit overall. Unfortunately, at present around 25% of our customers are churning.

This report provides some insights into the drivers of customer churn, using data mining and interpretable ("white box") models. Decision makers in the business can use these insights to inform strategic and product decisions, with the aim of reducing future churn, both for existing customers and potential new customers.

This report also provides an overview of a model optimised for predictive accuracy ("black box"). Such models can be used to derive customer lists for targeted retention campaigns.

High level findings include:

- Data mining highlights some clear drivers of churn, including being a senior citizen, being on a monthly contract and having shorter tenure.
- Machine learning models help us to identify the relative importance of these factors and control for correlation between them. In particular, the white box models indicate most of our churn is driven by senior citizens, with features such as monthly contracts and tenure featuring to a lesser extent.
- The black box model was able to provide a small uplift in predictive accuracy at the expense of interpretability, and higher precision was achieved at the expense of lower recall.

2. Data

The data used for this report consists of a single dataset with 7,032 records in which customers are flagged as either churn ("Churn") or not ("Retain"). This is the target variable of interest for both data mining and modelling.

Most of the data consists of discrete categorical features, with the exception of tenure and monthly/total charges. Customer number is included in the dataset and can be used to produce

customer lists for retention programs, but has been removed from the data used for the analysis. There are no missing values in the data.

The following plots provide some insights based on data mining the raw data.

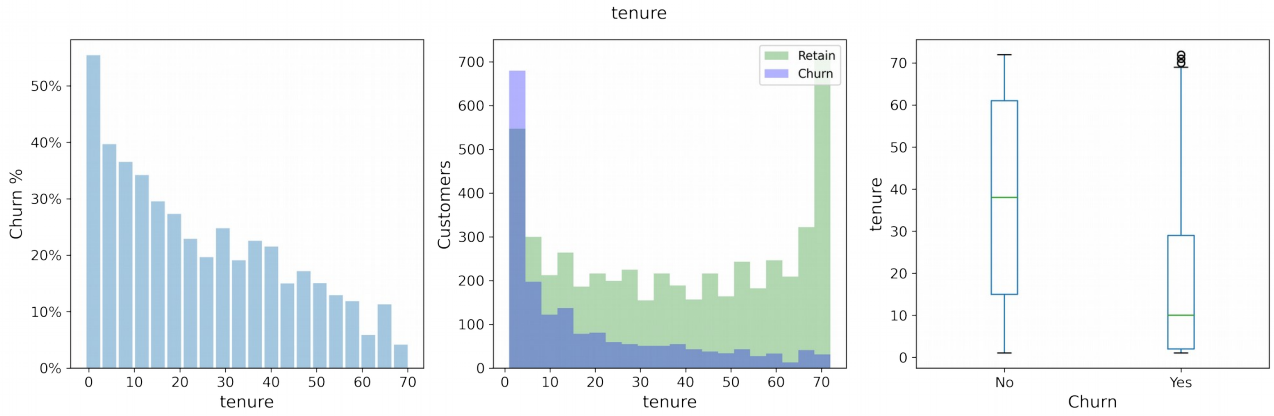


Figure 1: Churn by Tenure

The first panel shows the churn rate, the second shows customer numbers and the third shows a boxplot. Churn is significantly higher for customers with less tenure, and tends to reduce in a fairly linear fashion after the initial few months. While this feature is significant, it is of lesser value when considering potential new customers – we can't go looking for people who will ultimately have high tenure. For our existing portfolio, we may want to put greater focus on building the relaitonship with our new customers.

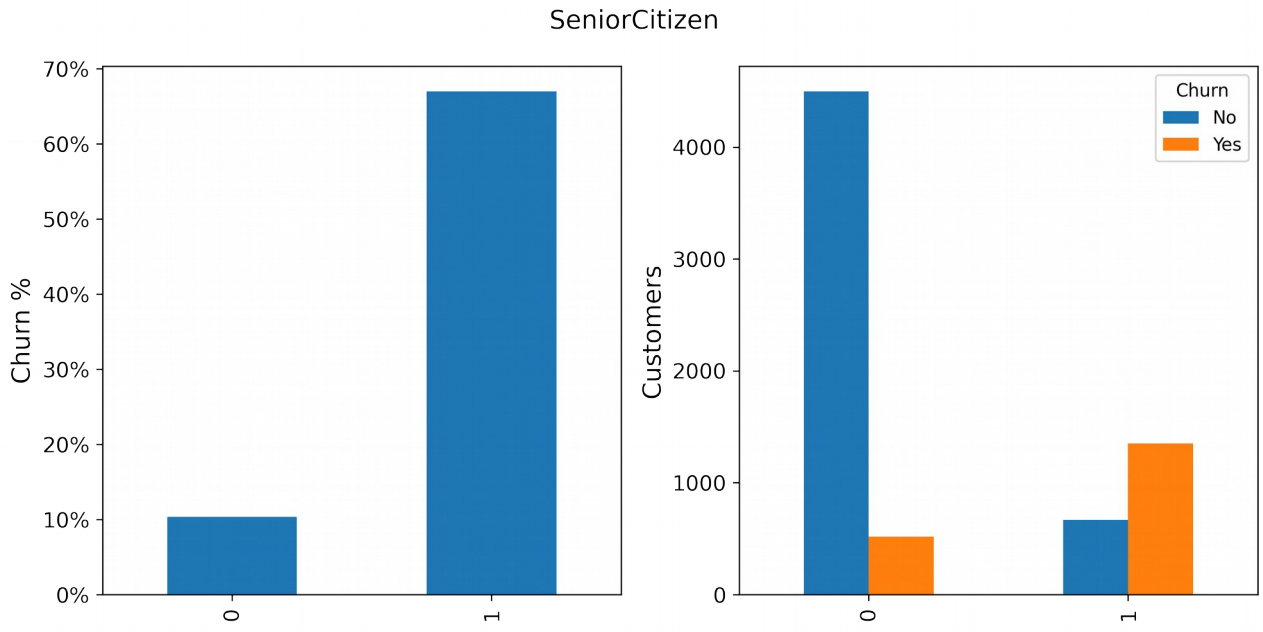


Figure 2: Churn by Senior Citizen

The first panel shows the churn rate, the second shows customer numbers. Churn is significantly higher for senior citizens, with a churn rate of two thirds. There is a clear opportunity to reduce churn by targeting this segment of our portfolio.

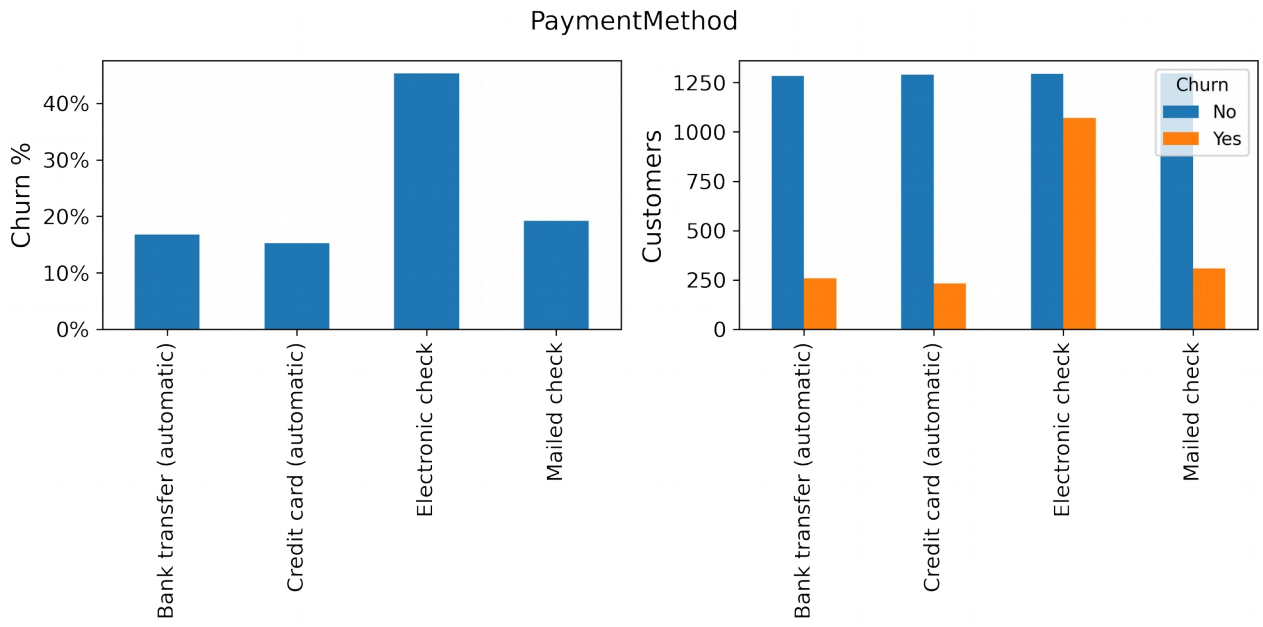


Figure 3: Churn by Payment Method

Churn is significantly higher for those paying by Electronic Check, at around 50% compared to around 15-20% for other payment methods. We may want to promote automatic billing to avoid a constant reminder of the amount being paid, which may prompt our customers to price shop ("out of sight, out of mind").

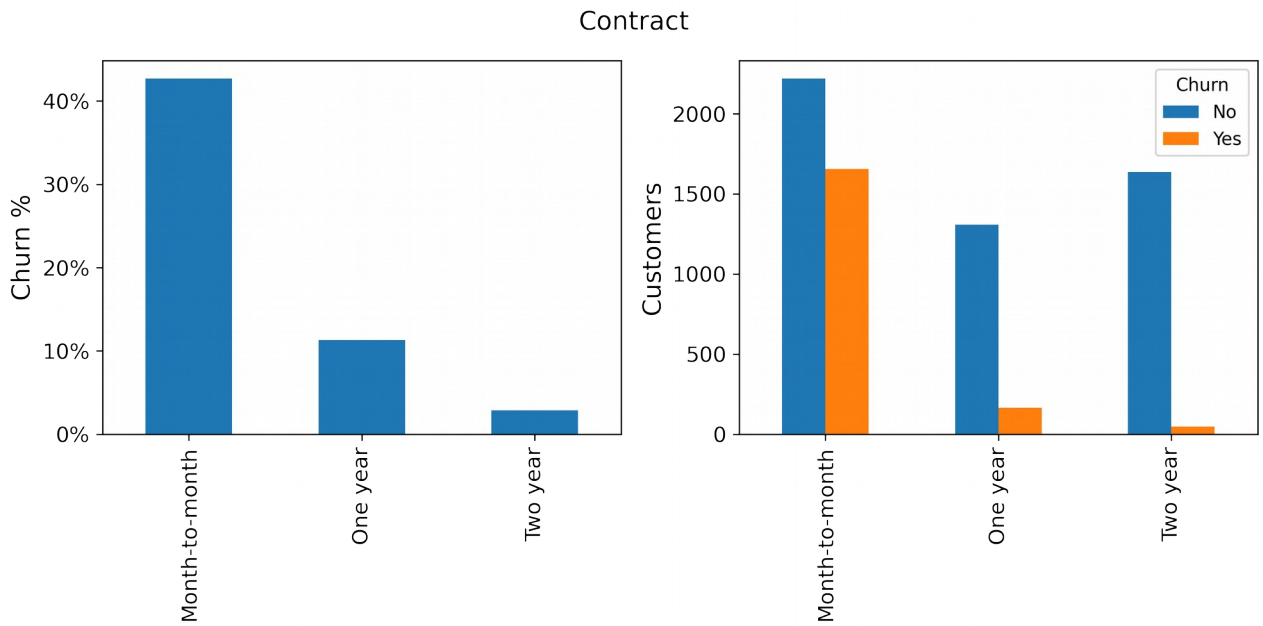


Figure 4: Churn by Contract

The impact of contract on churn is very clear, with month-to-month customers having much greater churn. To the extent that we can convince customers to "lock in", their propensity to churn will be much less. Any changes to contractual features such as an early termination fee, would need to consider the possible impact on churn. It's possible that we may be able to further reduce churn by increasing such penalties, but also possible that less customers would "lock in" given a higher penalty.

Figure 5: Churn by Online Security

There are a number of features in the data that relate to the products held by the customer. Online security is one such example, which has a clear relationship with churn (those with online security are less likely to churn), possibly reflecting brand loyalty, risk aversion or wealth characteristics of our customers.

Most of the product features include three levels, “Yes”, “No” and an additional level such as “No internet service”. This last level reflects information captured in another dedicated feature, and has been grouped with “No” in the models. This makes the model results more interpretable, ensuring the model does not rely on this feature unless it is due to online security in particular.

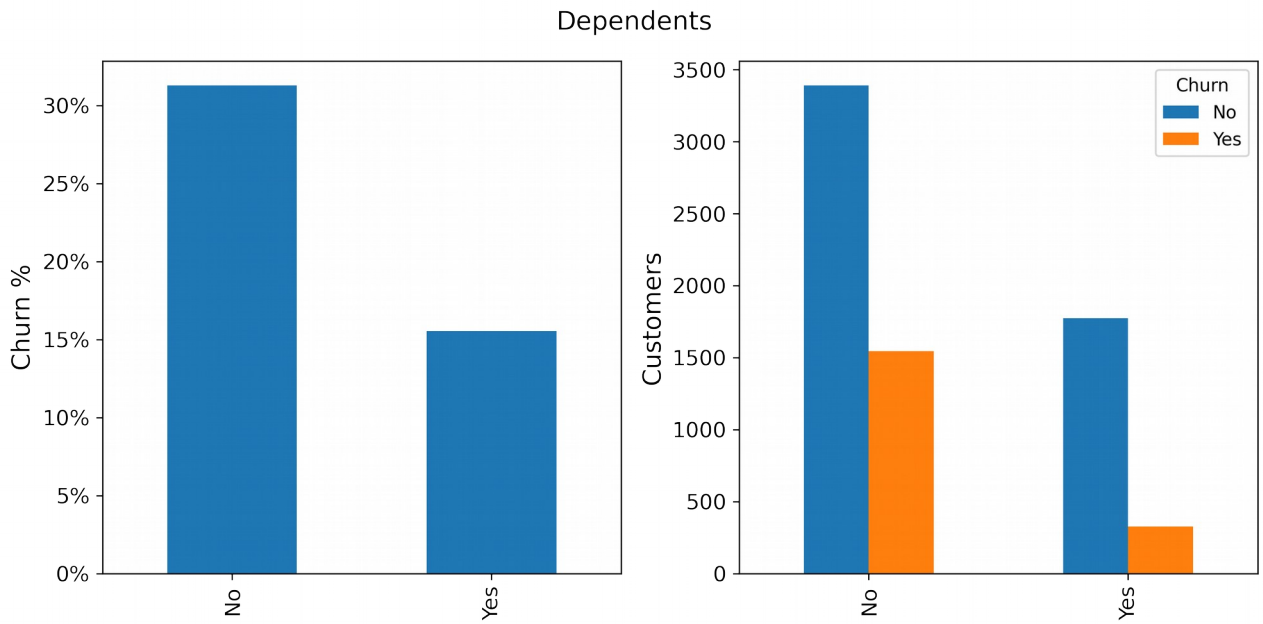


Figure 6: Churn by Dependents

Similar to the product features above, features such as Dependents and Partner reflect the family structure of our customers and have a clear relationship with churn.

3. Modelling

3.1 Methodology

Modelling was undertaken using the sklearn framework within Python. 80% of the data was used to train the models, with 20% of the data being used as a hold out set to be used at the end of the modelling exercise to verify that the final models were likely to generalise well and had not overfit the training data significantly. In splitting the train/test sets, stratification based on the churn column was used to ensure the classes were balanced.

For the “white box” interpretable models, Decision Trees (CART) were employed. The models were evaluated using a confusion matrix and a ROC AUC curve. The aim of this exercise was to identify insights from nodes in the tree which contain at least 50 customers, with at least 75% in the predicted class (corresponding to a gini of under 0.375). This approach ensures the insights found are robust and can have a measurable impact on business metrics.

The “black box” accuracy models were primarly based on ensemble tree methods such as Random Forest. The models were evaluated using 5-fold cross validation with learning curves, confusion matrices, ROC AUC curves and feature importance examined. The aim of this exercise was to obtain a precison of 80% and recall of 60% on the training data under cross validation. In general, F1 was used as a scoring metric when training the models, blending the precision/recall targets, although precision, ROC AUC and accuracy were also considered.

A genuine attempt was made to produce a model that would not only meet the target metrics but also generalise well to unseen data. It is possible, for example, to “cherry pick” a random state that meets the criteria, as shown on an early random forest model in Figure 7. A similar approach can be taken to tweak hyperparameters to overfit to the training data, even with cross validation in place.

Hyperparameters were selected using grid search or random search with 10-fold cross validation. For the white box models, min_samples_leaf was set to 50 to match the evaluation critiera, and grid search employed thereafter. For the black box models, grid search was used for logistic regression and random search (generally over 100 combinations) for the ensemble methods. An additional random search with a narrower range was applied in some cases after first identifying a promising candidate.

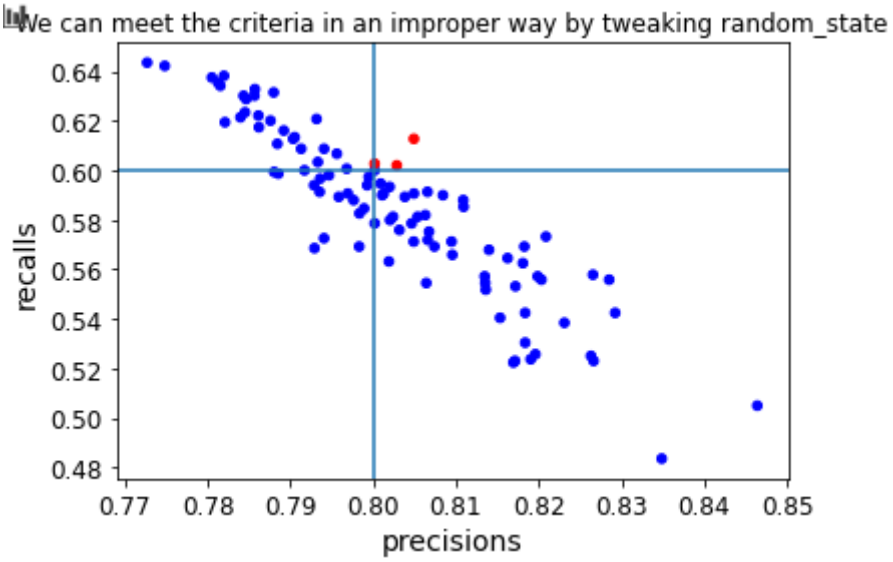


Figure 7: Tweaking random seeds to achieve targets

This model can achieve a cross validated precision of 80% and recall of 60% (red points), but probably won't generalise so well.

3.2 Data Preprocessing

Data transformation and feature selection was applied prior to modelling. In general, machine learning algorithms prefer numeric data, so fields encoded as text were encoded as numeric instead. One hot encoding was applied to convert columns with multiple levels into several binary columns. As mentioned previously, redundant levels in product columns were removed.

Total Spend was dropped from the analysis as it is correlated with Tenure and Monthly Spend. Tenure and Monthly spend were considered both as the original numeric variables, and separately as categorical variables. For tenure, the bands used were 1 month, 2-5 months, 6-12 months, 13-24 months, 25-36 months, 37-48 months, 49-60 months and over 60 months. For monthly spend, the bands used were low (less than $25), medium ($25-60) and high (>$65) (see Figure 8).

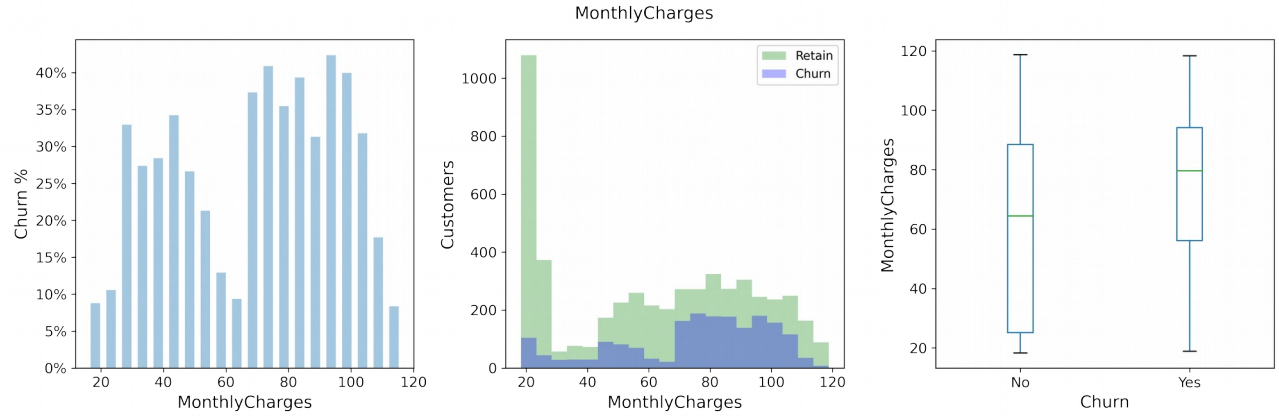


Figure 8: Churn by Monthly Charges

Many customers have low monthly charges, and these customers have a low churn rate (around 10%). There is another group of customers with moderate charges (around $50) with moderate churn (around 25%). A significant bulk of customers have higher monthly charges (around $70-110) and these customers have the highest churn (around 35%).

A few features were added to capture potential relationships between columns:

- Family structure – combining Senior Citizen, Partner and Dependents
- Risk averse – having any of Online Security, Online Backup, Device Protection or Tech Support products
- Current less historical monthly charges – comparing Total Charges / Tenure to Monthly Charges

The black box models were fit with and without these additional features.

3.3 White Box Models

The best decision tree, based on the methodology outlined in section 3.1, has depth 6. Figure 9 shows a smaller tree, which is easier to discuss and has essentially the same insights:

- The first split is based on senior citizen, with the model predicting that we'll retain all customers who are not senior citizens. This finding is consistent even for the best tree; only a single leaf predicts churn, for 50 of 74 customers who have monthly contracts, Fiber optic internet and are in their first month of tenure.
- Seniors on 1 or 2 year contracts are likely to be retained, although the gini is above our target threshold at this point. Even with the best tree, most leaves fail to meet the purity threshold, with the exception of seniors on 2 year contracts with automatic credit card payments, who are particularly likely to be retained.
- Seniors on monthly contracts are highly likely to churn. This finding is consistent even for the best tree, where only a single leaf predicts retain, for 31 of 53 customers who have long tenure and payment methods other than electronic check.

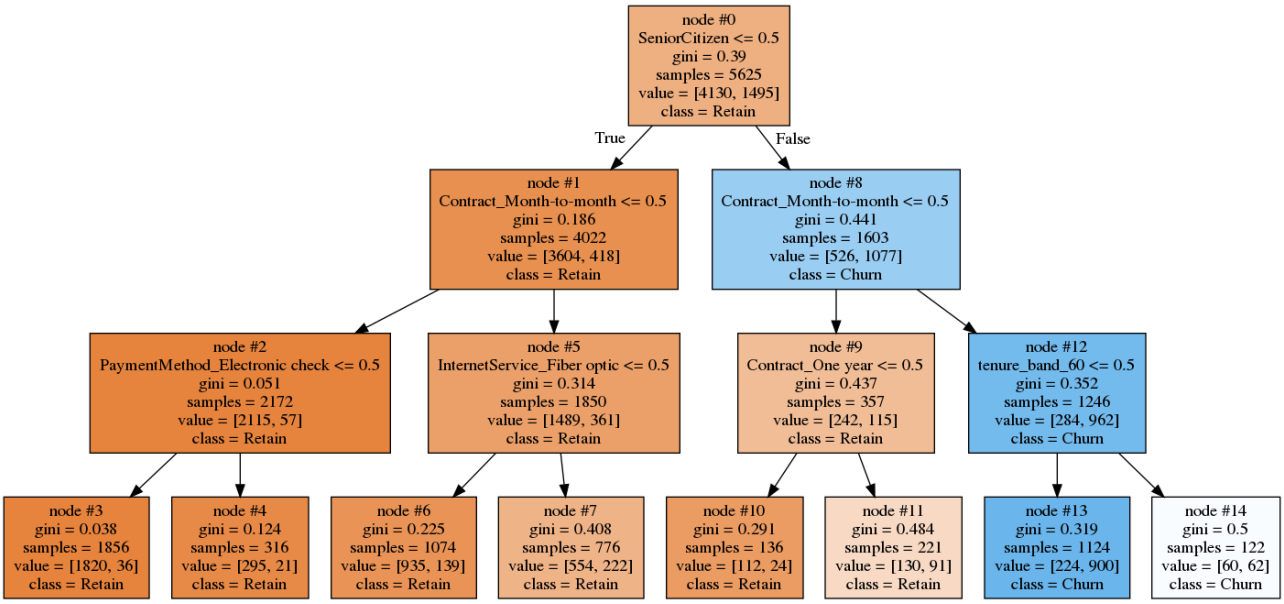
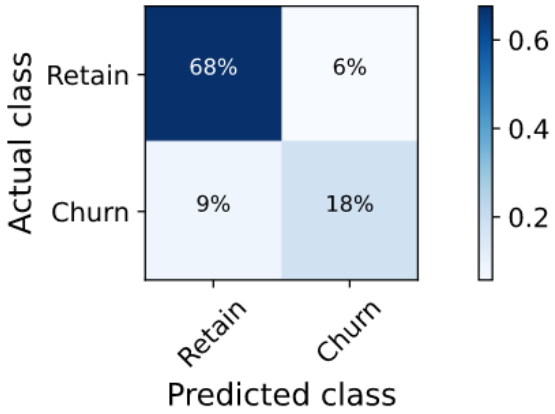


Figure 9: Small Tree

(max_depth=3, min_samples_leaf=50)

Figures 10-12 show the evaluation metrics (using cross validation on the training set) for the best tree (with max_depth 6). The small tree shown previously has slightly worse metrics overall (AUC 88%, Precision 77%, Recall 63%), but both trees demonstrate fairly good predictive performance.



Precision=76% (% Predicted Churn which are Actual Churn)
Recall=67% (% Actual Churn which are Predicted Churn)
F1=71% (Harmonic mean of Precision and Recall)
Accuracy=86% (% Predicted which match Actual)

Figure 10: Confusion Matrix - Decision Tree

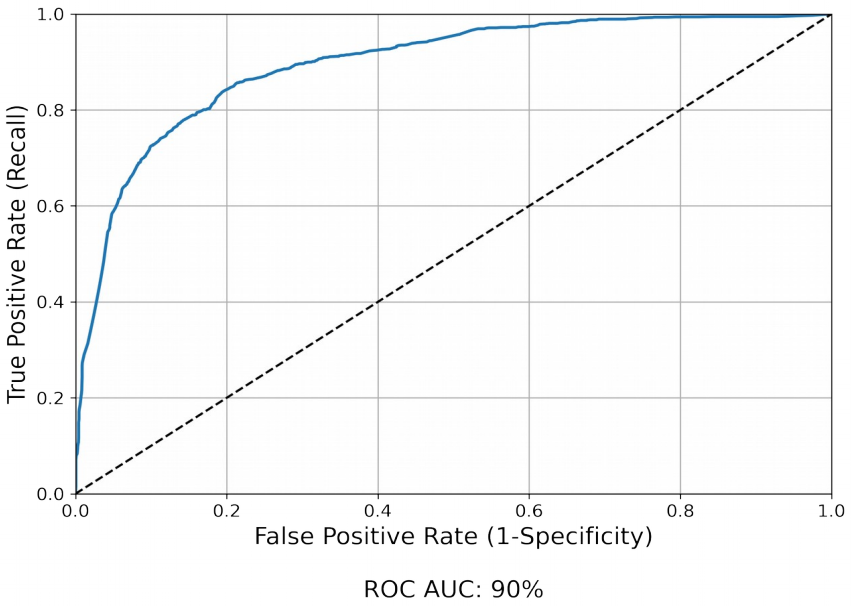


Figure 11: ROC AUC Curve - Decision Tree

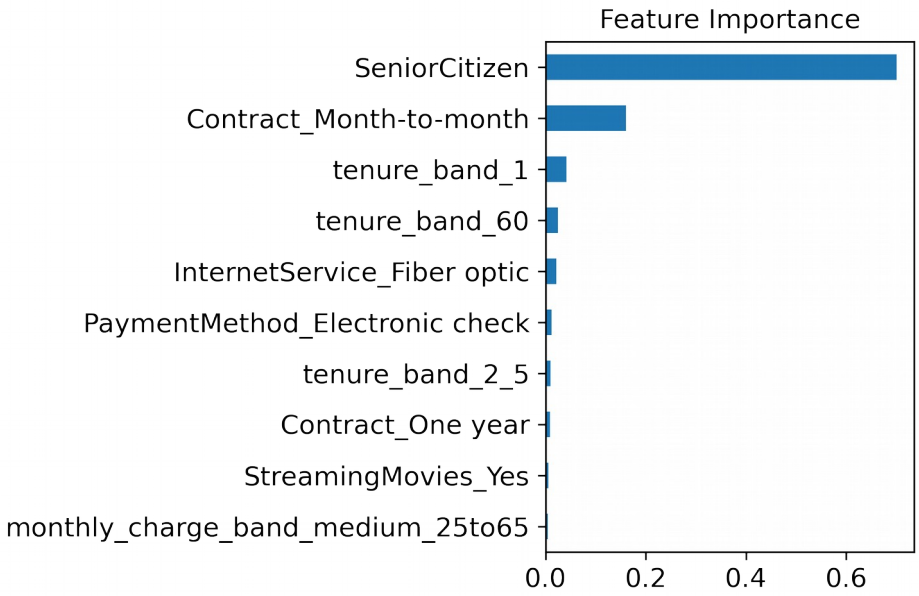
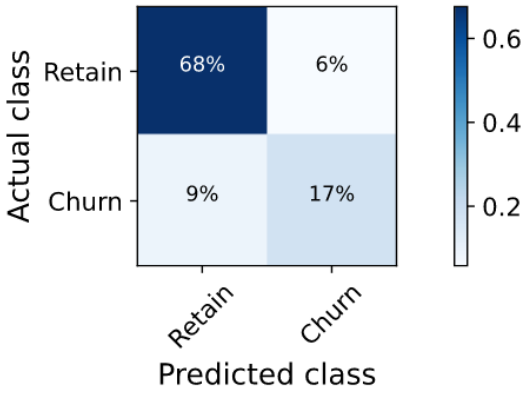


Figure 12: Feature Importance - Decision Tree

The feature importances reflect the discussion above; senior citizen is the most significant predictor by a long way, with monthly contracts and short/long tenure providing some further indication of whether customers are likely to churn.

Finally, Figure 13 shows the confusion matrix of the best model applied to the test set. Performance is comparable to that of the training set.



Precision=75% (% Predicted Churn which are Actual Churn)
Recall=66% (% Actual Churn which are Predicted Churn)
F1=70% (Harmonic mean of Precision and Recall)
Accuracy=85% (% Predicted which match Actual)

Figure 13: Confusion Matrix on Test Set – Decision Tree

3.4 Black Box Models

A range of models were considered. There was no particular stand-out model, with many models achieving similar performance, with precision close to 80% given recall of at least 60%. This report documents one such model as well as a few insights from another experiment.

3.4.1 Random Forest Blend

A random forest model was trained with hyperparameters selected based on F1 as a scoring metric[1]. Compared to the target levels, this model had higher recall but lower precision. A second random forest was then trained with hyperparameters selected based on precision as a scoring metric. This model was shallower and used less estimators, resulting in higher precision at the expense of lower recall. These two models were then blended using a soft voting classifier, with 80% weight given to the precision model. The learning curves from this final model are shown in Figure 14.

1 For this model, the version of the data with categorical banding of tenure and monthly spend was used.

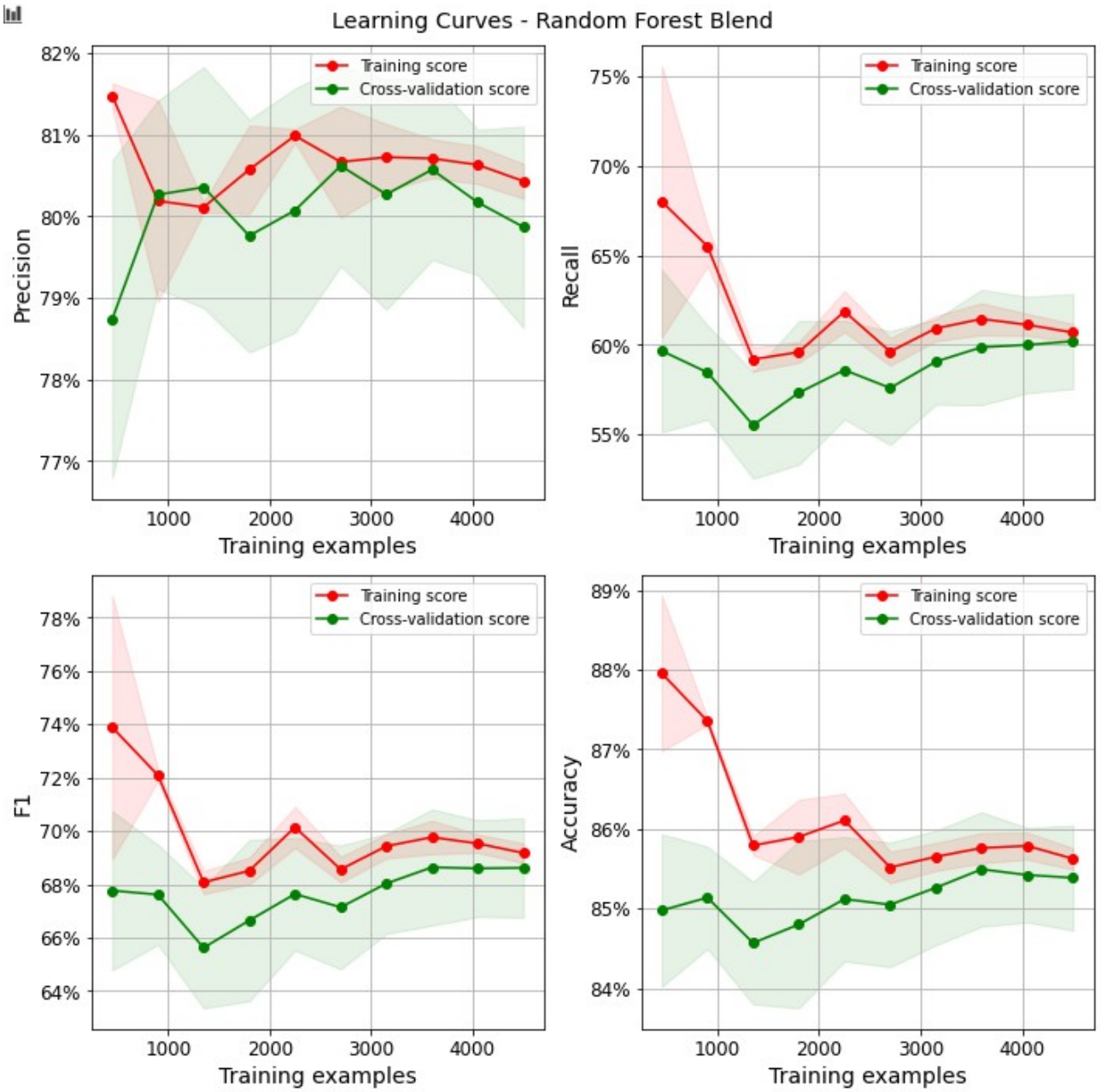
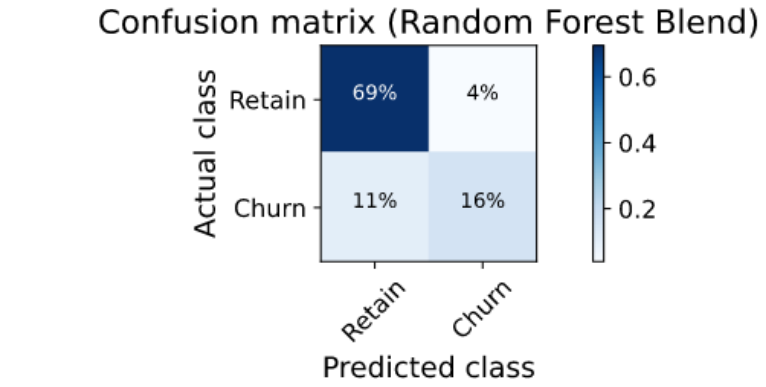


Figure 14: Learning Curves for Random Forest Blend

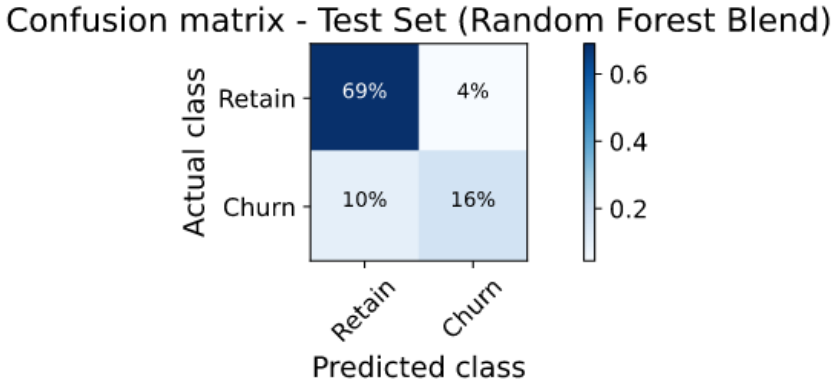
The blended model produces cross validated scores for precision and recall that are at the target range. The higher weight given to the precision model (80%), in addition to the natural weight given to precision when using F1 scoring, highlights that the precision target was generally the more difficult target to achieve.

Figures 15 and 16 show the confusion matrices and feature importances for the precision and F1 models respectively. Performance on the test set is comparable to that on the training set under cross validation, with a small increase in recall and small drop in precision.



Precision=80% (% Predicted Churn which are Actual Churn)
Recall=60% (% Actual Churn which are Predicted Churn)
F1=69% (Harmonic mean of Precision and Recall)
Accuracy=85% (% Predicted which match Actual)

Figure 15: Confusion Matrix - Random Forest Blend



Precision=78% (% Predicted Churn which are Actual Churn)
Recall=61% (% Actual Churn which are Predicted Churn)
F1=69% (Harmonic mean of Precision and Recall)
Accuracy=85% (% Predicted which match Actual)

Figure 16: Confusion Matrix on Test Set - Random Forest Blend

3.4.2 Four Model Blend

The learning curves shown previously suggest that the models are able to be trained on smaller data to similar effect. Based on this, I used 80% of the training data to fit individual predictor models using Random Forest, Adaboost, Logistic Regression and Extremely Randomized Trees algorithms, and used the remaining 20% of the training data to fit a Logistic Regression to blend between the individual predictor models.

A simple blend was used, as well as a feature-weighted blend where predictor model results were crossed with some of the more predictive features. In theory, this would allow, for example, a model that was better at predicting results for non-seniors, to receive more weight than other models when predicting non-seniors. In practice, these blending approaches didn't significantly improve results, possibly due to the predictors being too similar in nature. Including more varied algorithms such as neural networks or SVM may have made this approach more fruitful.

4. Conclusion

Data mining identified a number of relevant features to help understand the drivers of churn. White box models were used to identify the relative importance of these, with senior citizens being the most significant feature, followed by monthly contracts and tenure. Black box models were used to provide a slight increase in accuracy, however the uplift was relatively minor and higher precision was generally obtained at the cost of lower recall. This allows us to be more targeted in our customer retention efforts and minimise the costs of a customer retention program, at the expense of the program reaching less potential churn customers.