Churn Prediction in the Telco industry

Clarence San

Code: https://github.com/csanry/machinelearning/blob/main/telco_churn_classification/CODE/telco-customer-churn.ipynb

Central Questions

What is the profile of our customers most likely to churn?

What should we look for to predict whether a customer is likely to churn or not?

Background and Context

Churn is particularly important for service-based businesses such as telco companies – knowing which accounts are likely to churn will allow companies to apply retention strategies effectively to maximize customer lifetime value.

This IBM dataset has information about customers and if they have churned. This project aims to predict churn given characteristics about the customer account.

4	6	8	1	7043
Demographic Variables	**Account Info Variables**	**Service-related Variables**	**Churn Indicator**	**Accounts**

[1] Churn is defined as customers who left their contract within the past month

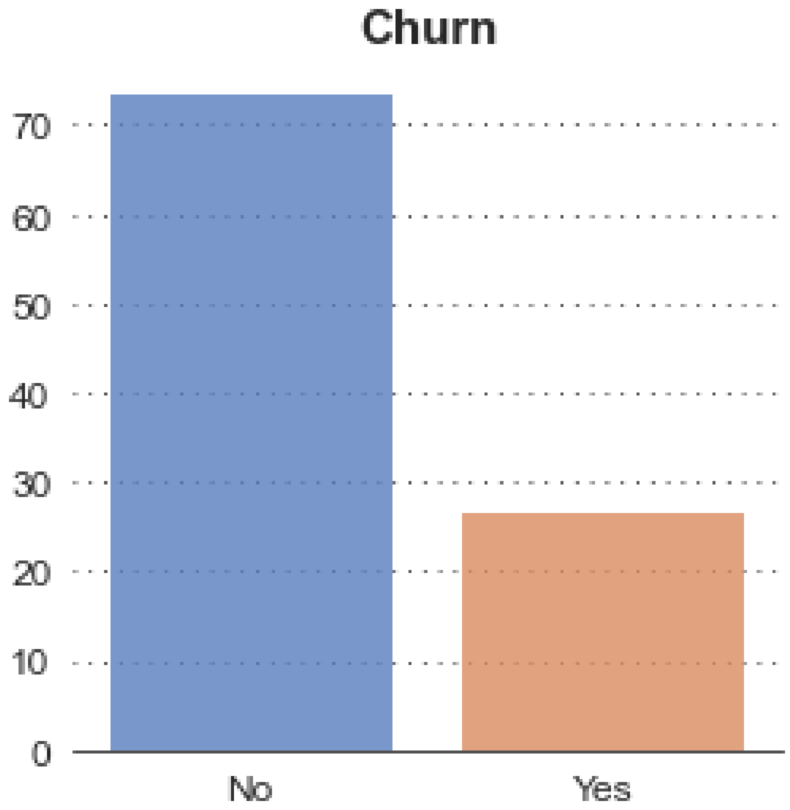
Exploratory Data Analysis

Process

Plotting distributions of numerical variables

Plotting counts of categorical variables

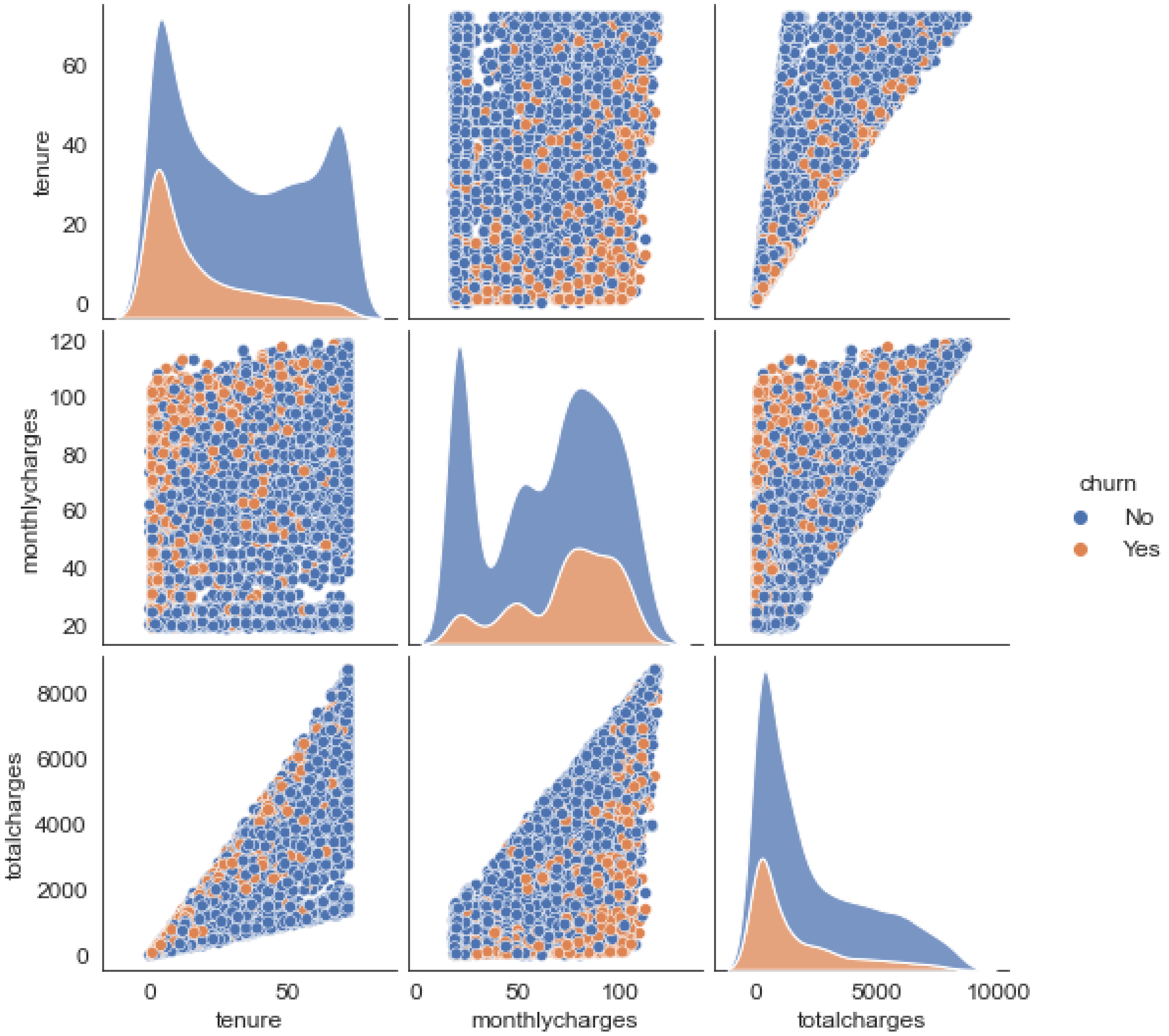
Churn Rates of Customers



Customers are on average 73% likely to churn

This provides a good baseline to evaluate model performance

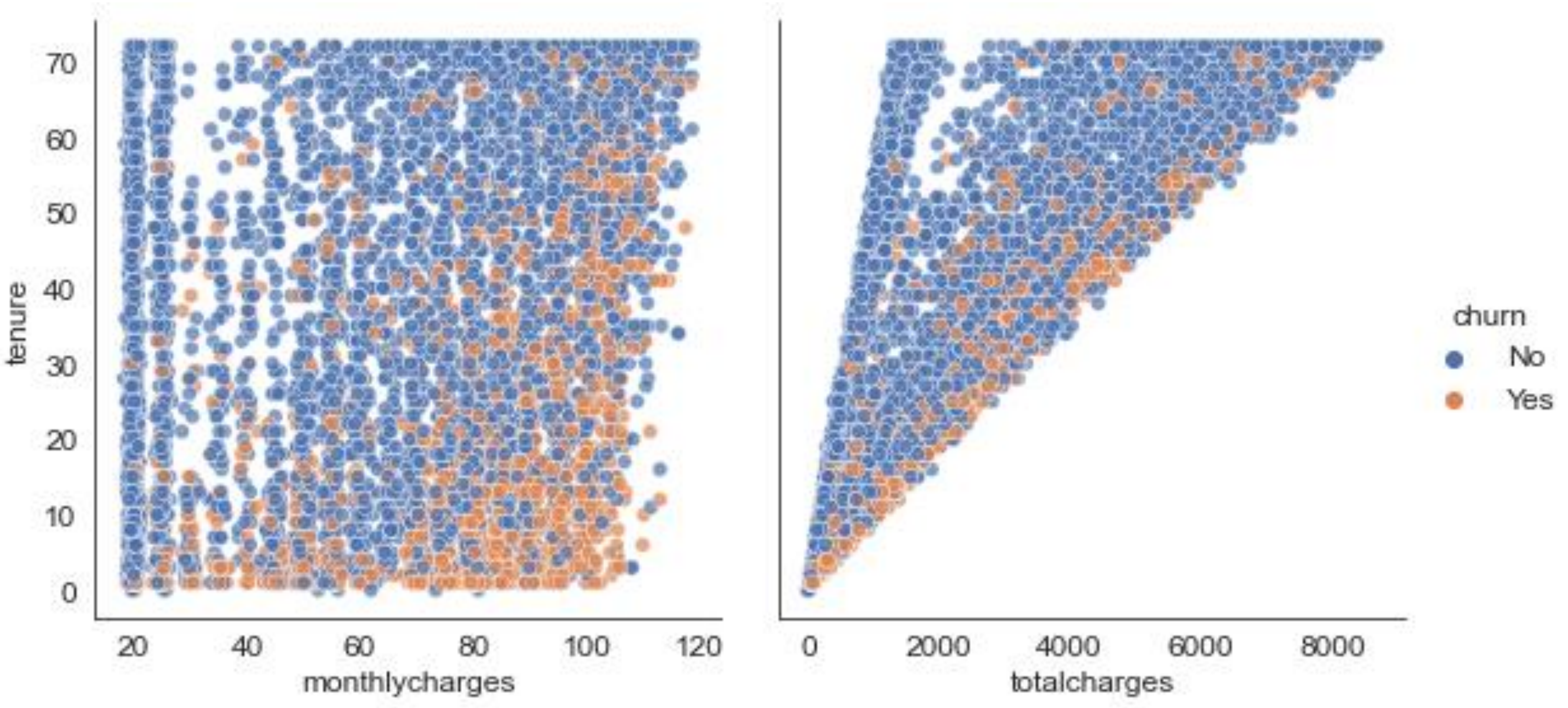
General Pairplot



There are 3 numerical columns – we can estimate their distributions using kernel density estimation

[1] Data points and kde distributions are split by churn result

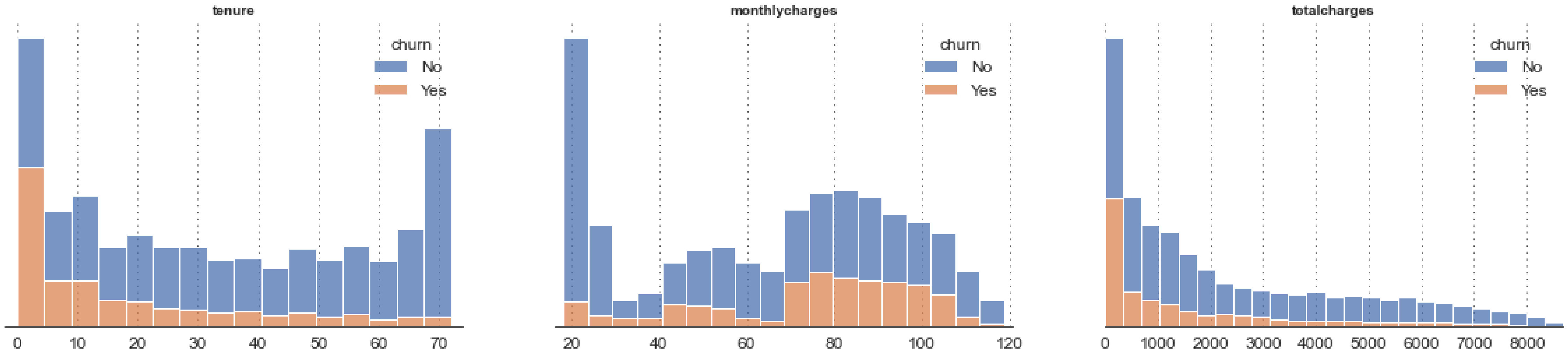
Facet monthlycharges and totalcharges by tenure



Observations

- Newer customers seem to be more likely to churn
- Customers with more expensive monthly plans are also likely to churn

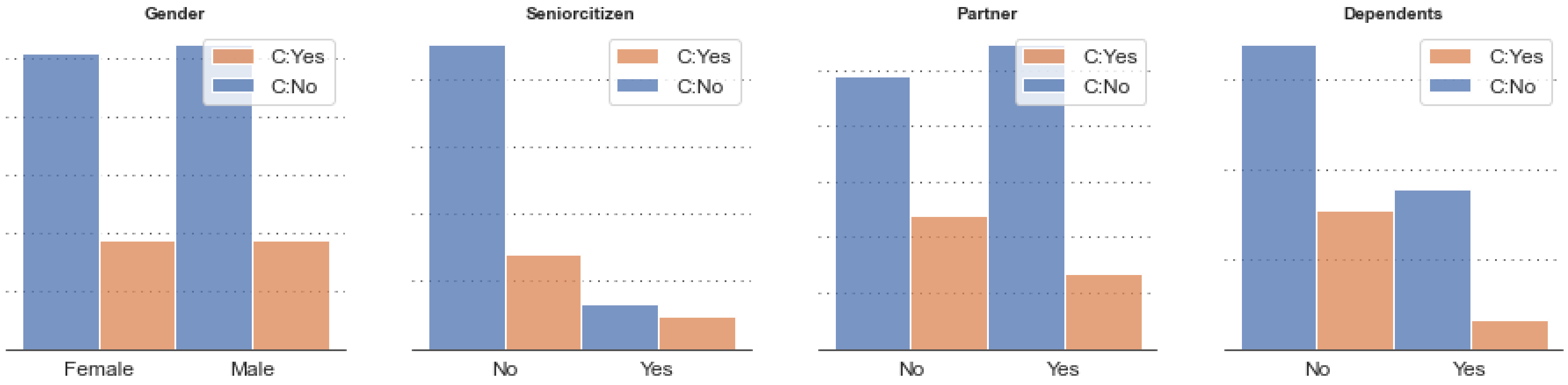
KDE plots of numerical features



Observations

- Churn rates higher at lower tenures (as well as for customers on monthly plans)
- The rate drops markedly from the 5 month period
- Lower charges are associated with lower rates of churn

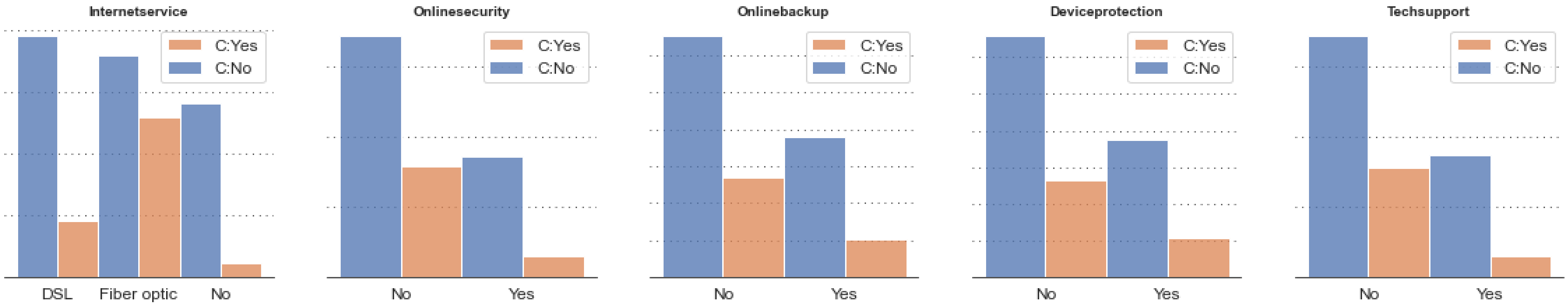
Demographic Breakdown



Observations

- Accounts with dependents tend to churn less
- There is little difference in churn between genders and individuals with partners

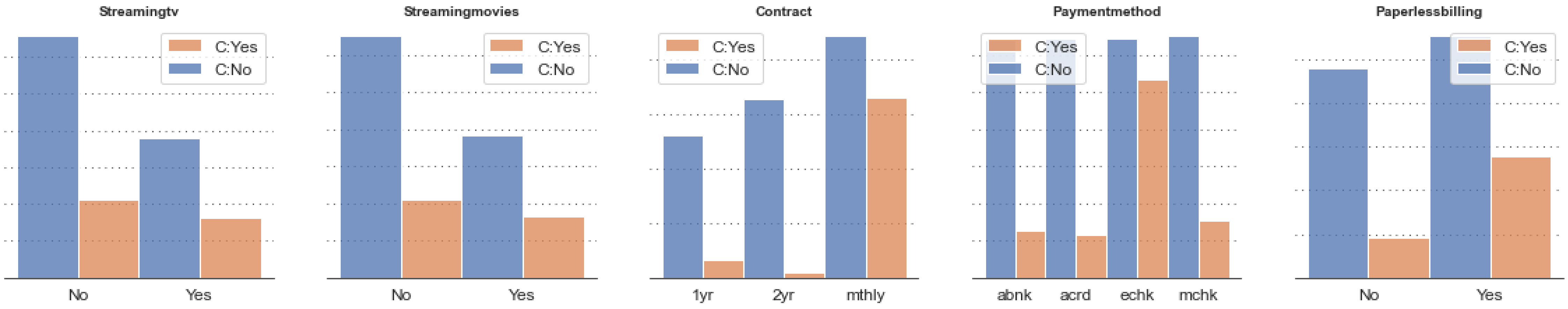
Services Breakdown



Observations

- Fiber optic subscriptions are more likely to churn
- Accounts with Internet-related services are less likely to churn

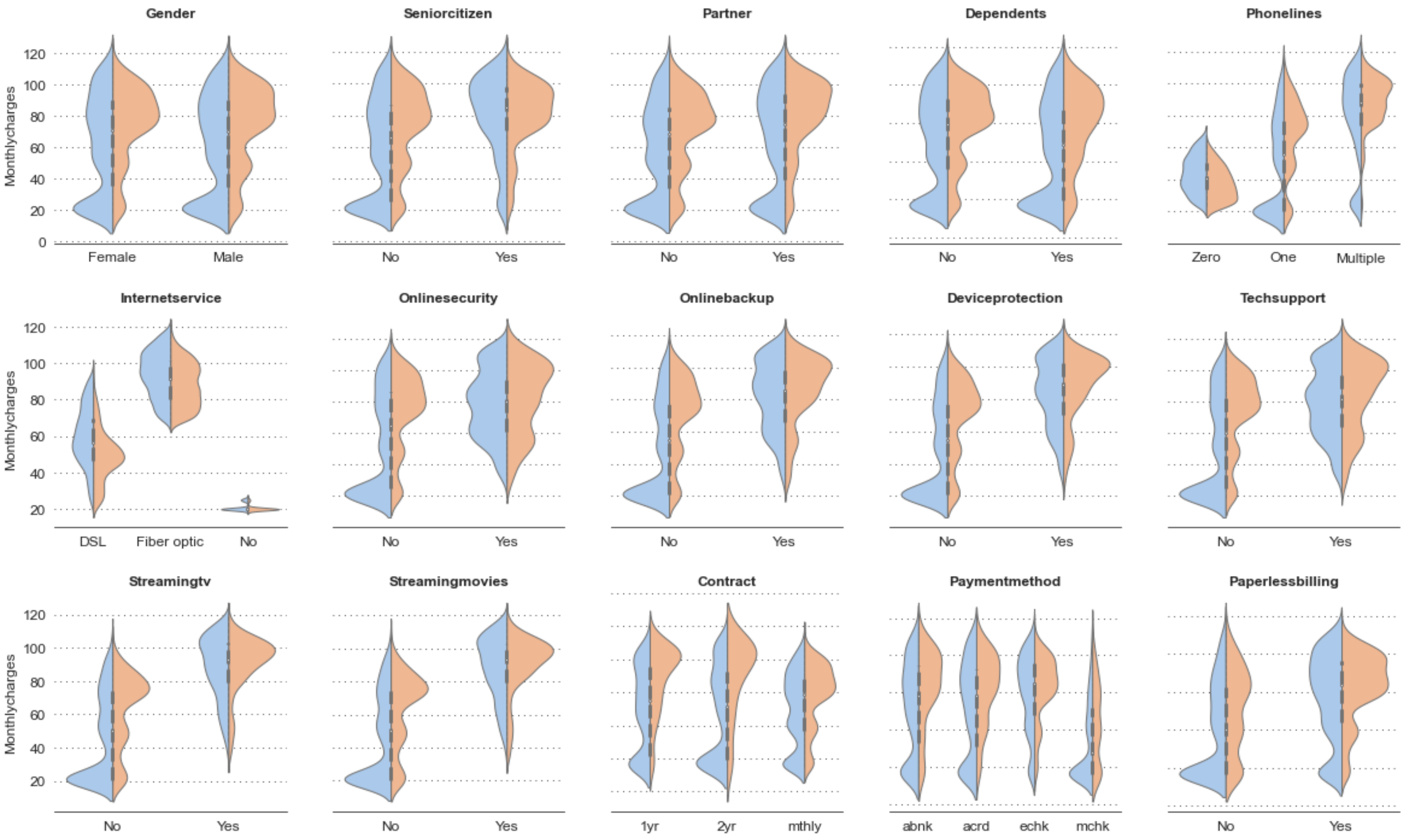
Account Breakdown



Observations

- Accounts without automated payments are more likely to churn
- There is a small increase in percentage churn for accounts with streaming services
- Monthly contracts are more likely to churn

Violin plots of Variables



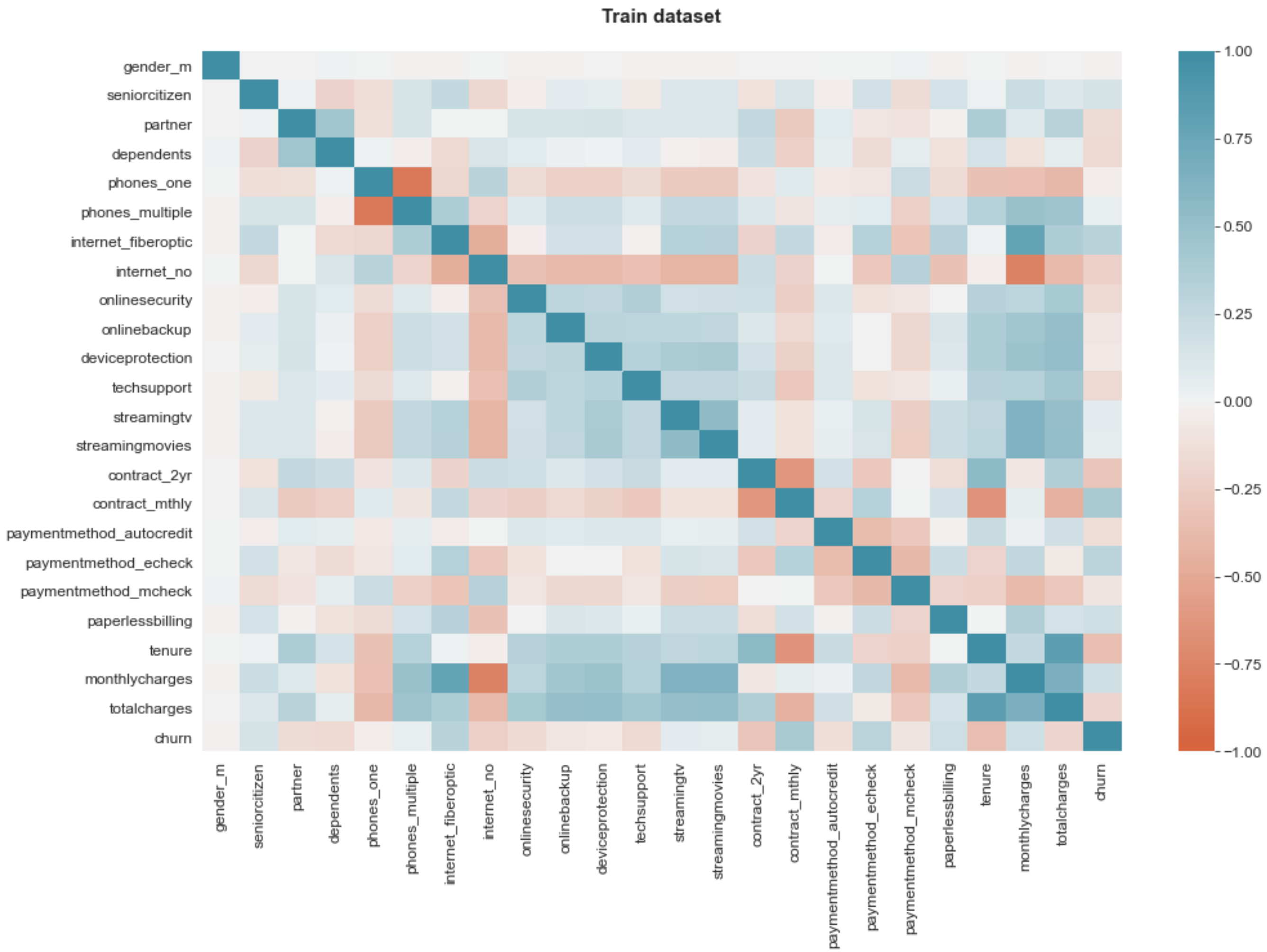
We can see that fiber optic subscriptions are more expensive than DSL, which might contribute to higher churn rates

Correlations

Process

Indepth look into the relationships among variables

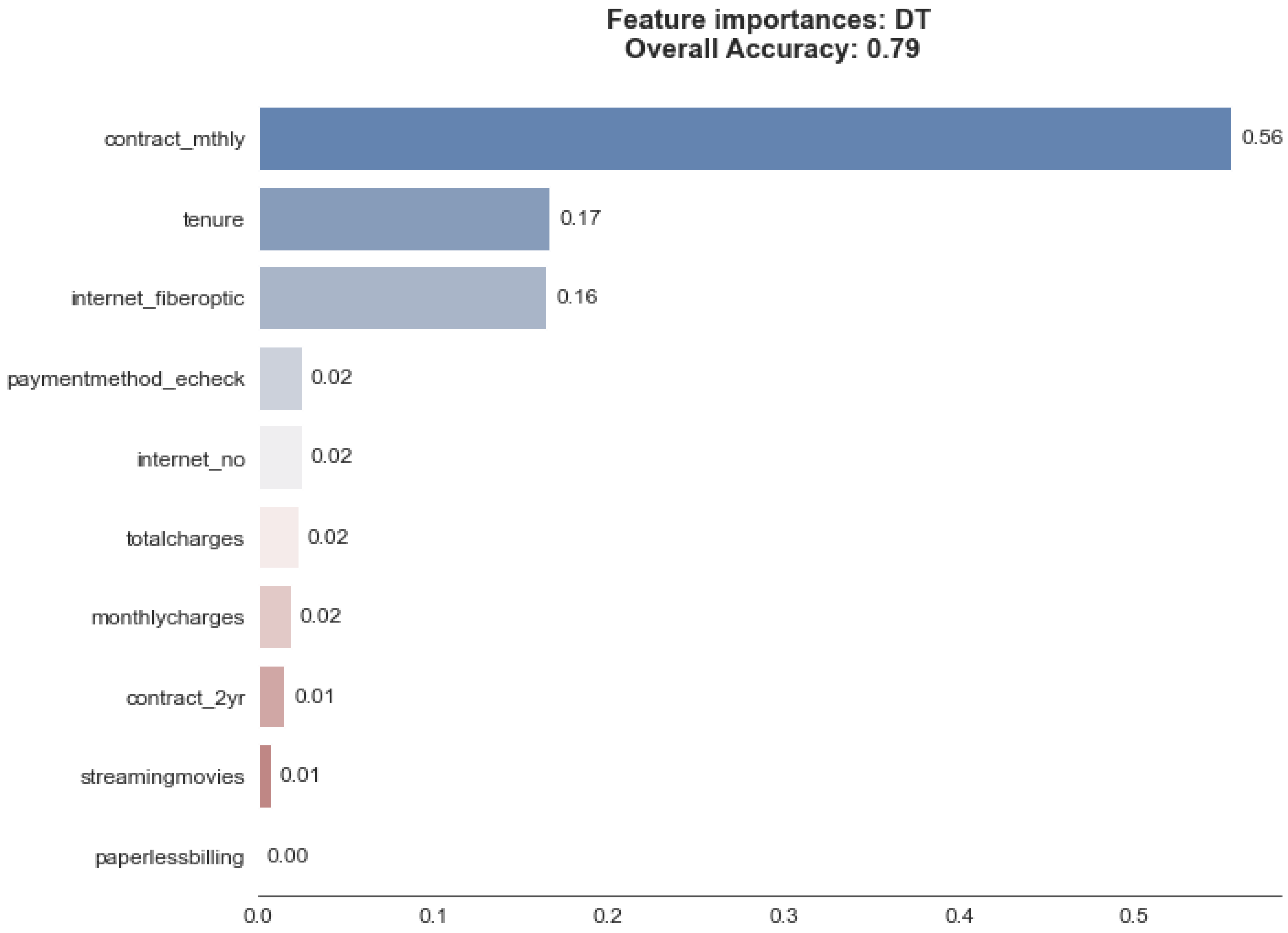
Retention Rate by Cohort



Notable associations to churn include age, fiber optic users, e-payments, and contracts.

[1] Pairwise correlations matrix across all variables

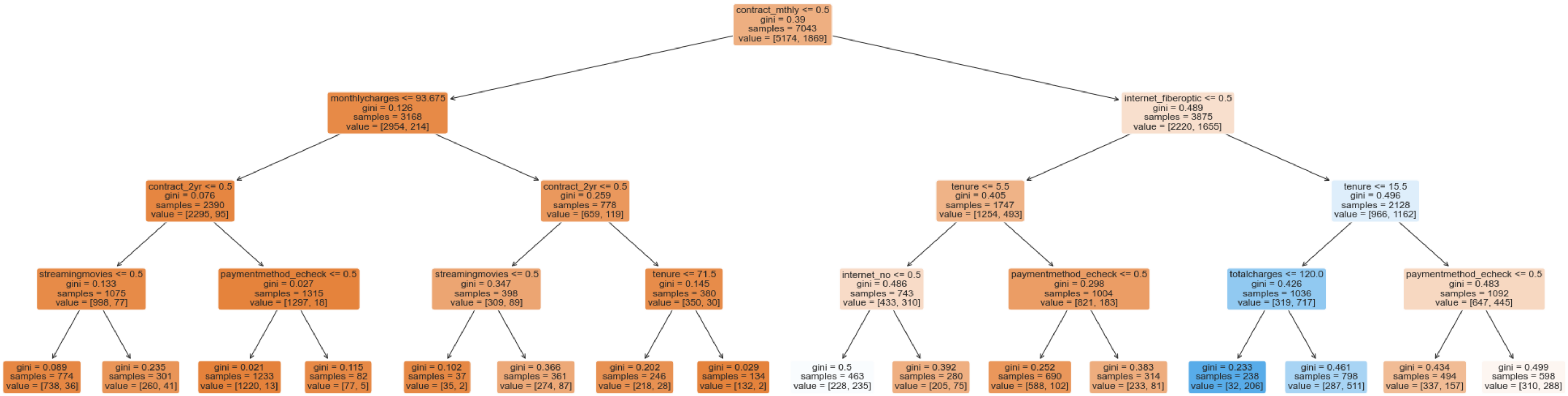
Importance Scores



Contract type, tenure and internet type have higher importance scores when trained on an initial model

[1] Decision Tree Classifier trained on all features, with max depth set to 4 to control test size; overall accuracy of 0.79

Visualization of the Decision Boundaries in the Tree



[1] Decision Tree Classifier trained on all features, with max depth set to 4 to control test size; overall accuracy of 0.79

Modelling Results

Process

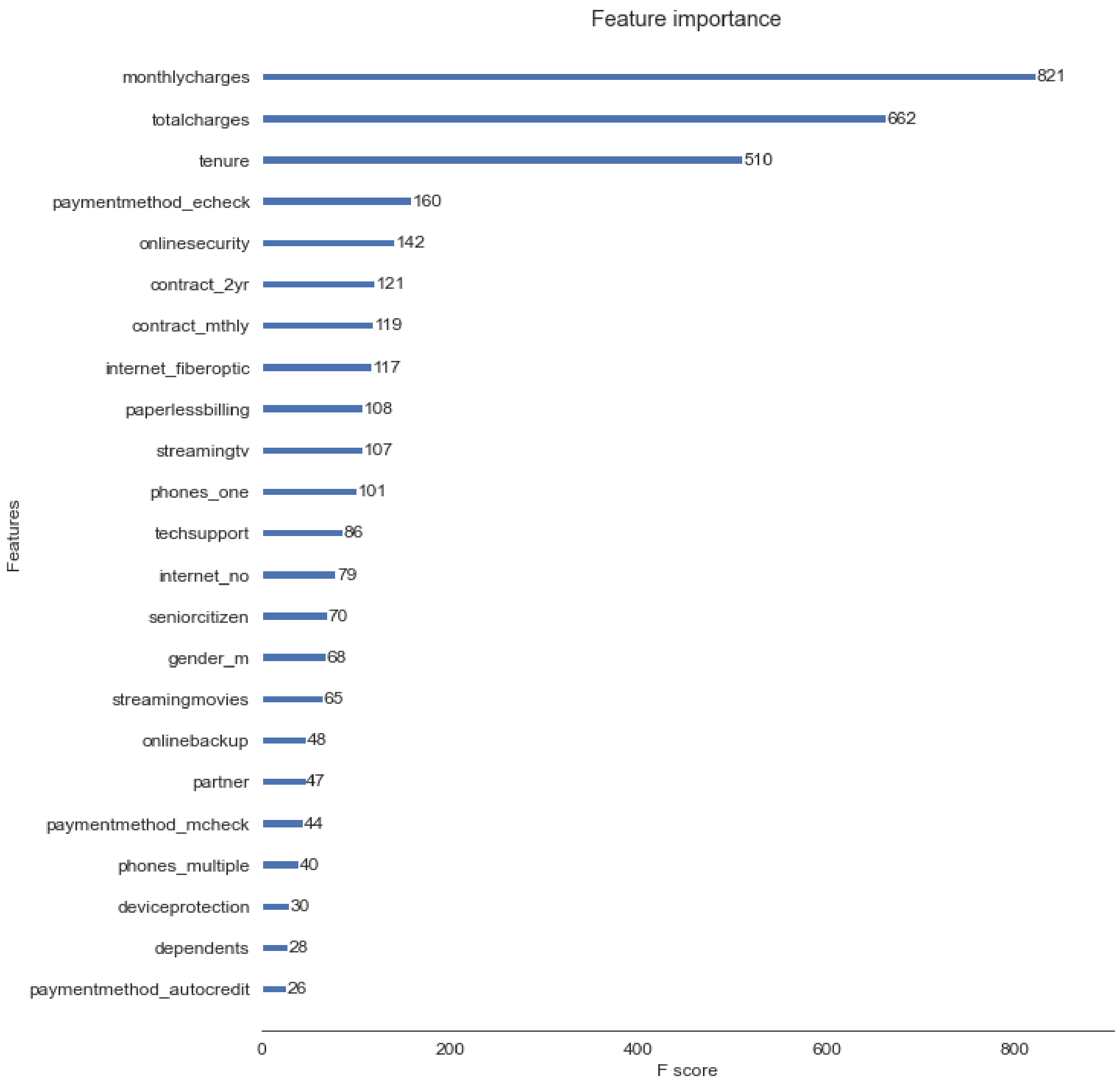
CART ensembles (Gradient Boosting and Random Forest) were tuned and validated

Potential Follow ups

As an additional step, we might want to consider dropping variables that are not important to the model in predicting churn

(Step is executed in code)

Feature Importance of the best performing model



The overall model has an accuracy of 78%, against a baseline of 73%

Monthly contracts, total charges, and tenure are the most important to churn prediction

Tuned Random Forest parameters
{'colsample_bynode': 0.6, 'learning_rate': 1.05, 'reg_lambda': 0.14, 'subsample': 0.79, 'objective': 'binary:logistic', 'base_score': 0.5, 'booster': 'gbtree', 'colsample_bylevel': 1, 'colsample_bytree': 1, 'gamma': 0, 'importance_type': 'gain', 'max_depth': 6, 'min_child_weight': 1, 'missing': nan, 'monotone_constraints': '()', 'n_estimators': 75, 'num_parallel_tree': 100, 'reg_alpha': 0.14, 'scale_pos_weight': 1, 'tree_method': 'exact', 'validate_parameters': 1, 'verbosity': 0}

Tuned Random Forest model validation
0.84 ROC-AUC, 0.78 Accuracy, 0.48 Recall, 0.66 Precision, 0.55 F1

[1] Random Forest was selected over a Gradient Boosted approach due to overall performance and ability to generalize to unseen data

Conclusion

Important features: fiber optic, streaming service, tenure
Not so important: gender, seniority, dependents

Gaining an understanding of the characteristics of churners is important for any company's retention strategy. A churn prediction model is also able to provide actionable insights and outputs to target potential churners

Next Steps

Investigate additional features that can be used to improve model’s predictive performance

Code: https://github.com/csanry/machinelearning/blob/main/telco_churn_classification/CODE/telco-customer-churn.ipynb