Predict, Understand and Reduce Customer Churn

EXECUTIVE SUMMARY

- All code is wrapped into a Python lib with a consolidated pipeline provided to run all the jobs (see “pipelines.html” or the code base for details)
- Three types of models (Logistic, Random Forest and XGBoost) are trained to predict customer churn using the given dataset, achieving:
  - 71% GINI
  - Capture of 50% churns within the top 20% “risky” customers

  The predictions can be used to identify ***already*** “risky” customers for intervention
- Driving factors of churn are scrutinised. It provides guidelines to establish and maintain a healthy customer loyalty/stickiness profile ***before*** they become “risky”
- Plan to estimate the models’ value using Cost Benefit Analysis is proposed
- Possible improvements of this work are proposed

Contents

1. **Exploratory Analysis (EPA) and Problem Statement**
2. Preprocessing
3. Model Training
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. Next Steps

The given dataset is explored to formulate the problem to solve: predict, understand and reduce churn

- The following items are explored
  - Distribution of the target to predict
  - Data types of features (Numerical and Categorical)
  - Distribution of features

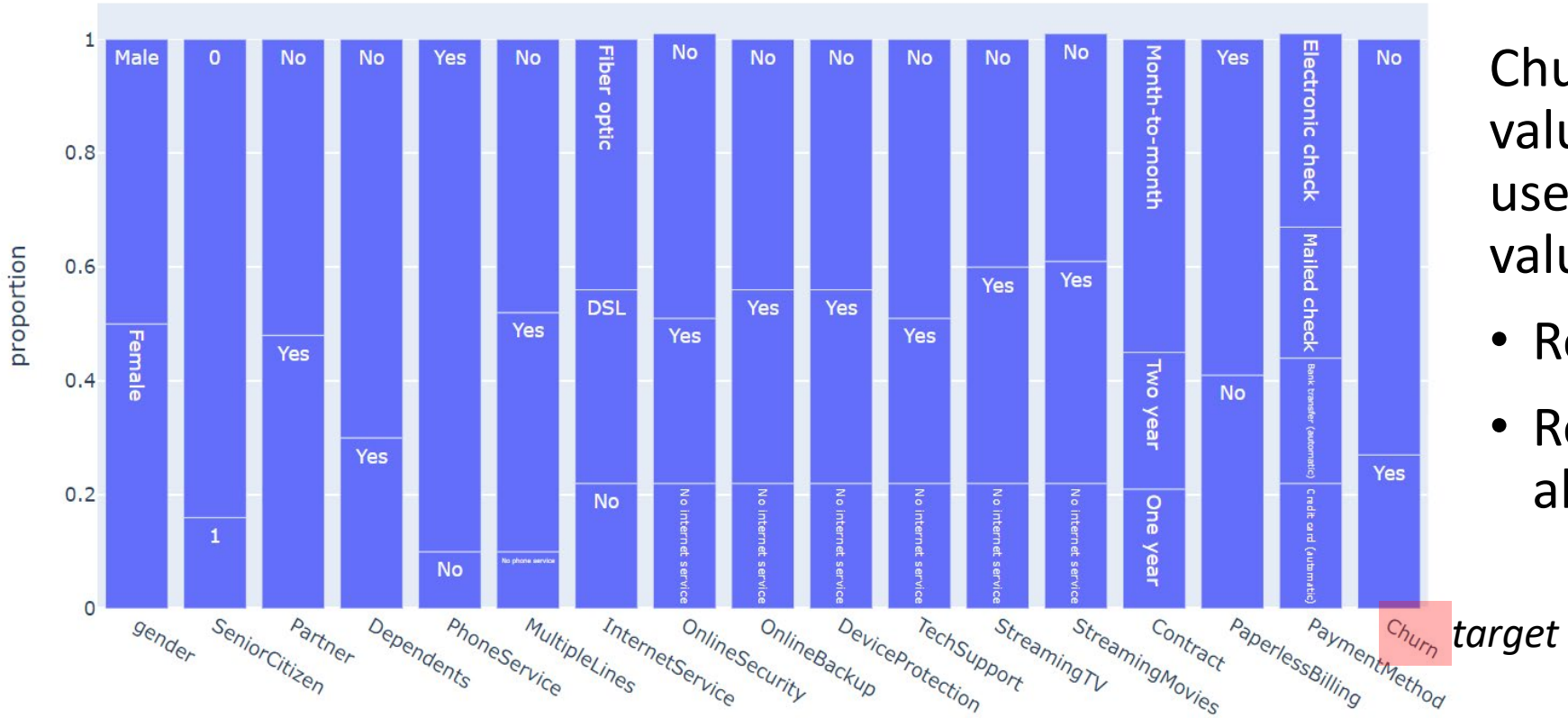
  Details can be found in the file “playground.html” or “playground.ipynb”

- The main scope is thus defined as
  - Predicting churn (a relatively balanced binary classification problem)
  - Clarifying the driving factors of churn
  - Providing suggestions to reduce churn

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. **Preprocessing**
3. Model Training
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. Next Steps

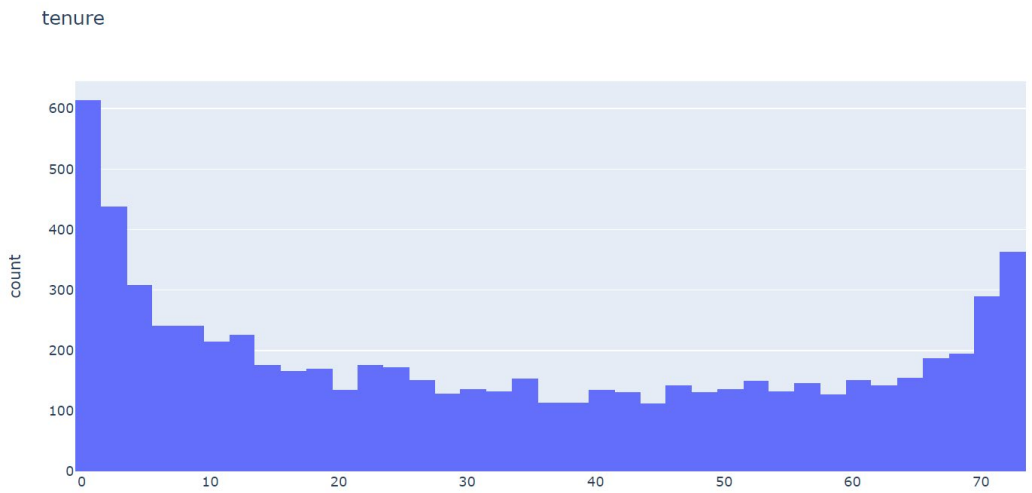
Categorical features are target encoded based on their distribution and the reasonable dataset size

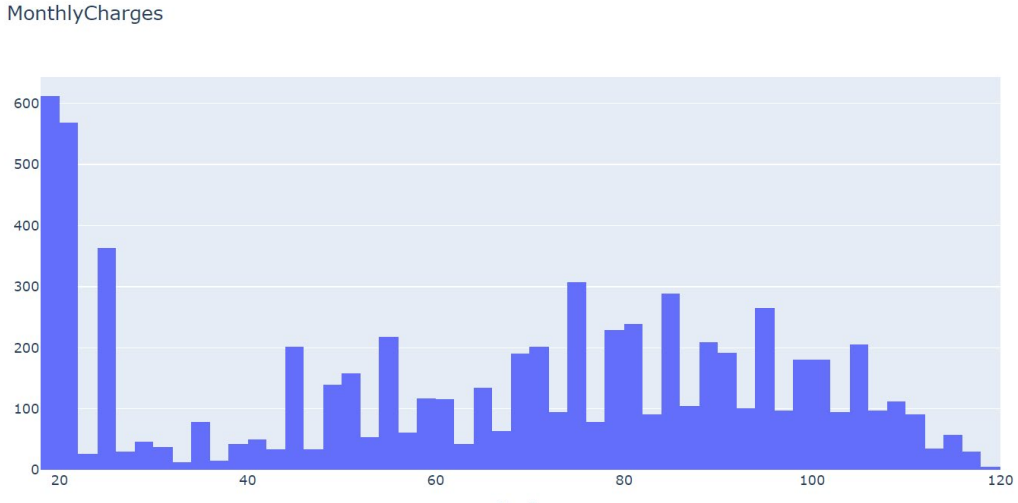


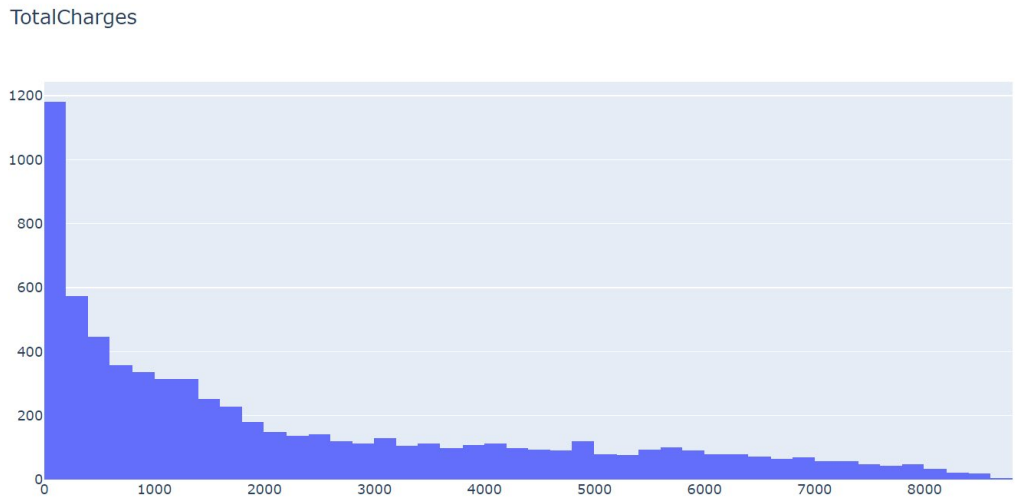
Churn rate of each unique value (group) for a feature is used to replace the original value given that:

- Reasonable sample size
- Relatively low cardinality of all categorical features

Numerical features are modelled as continuous features based on their distribution

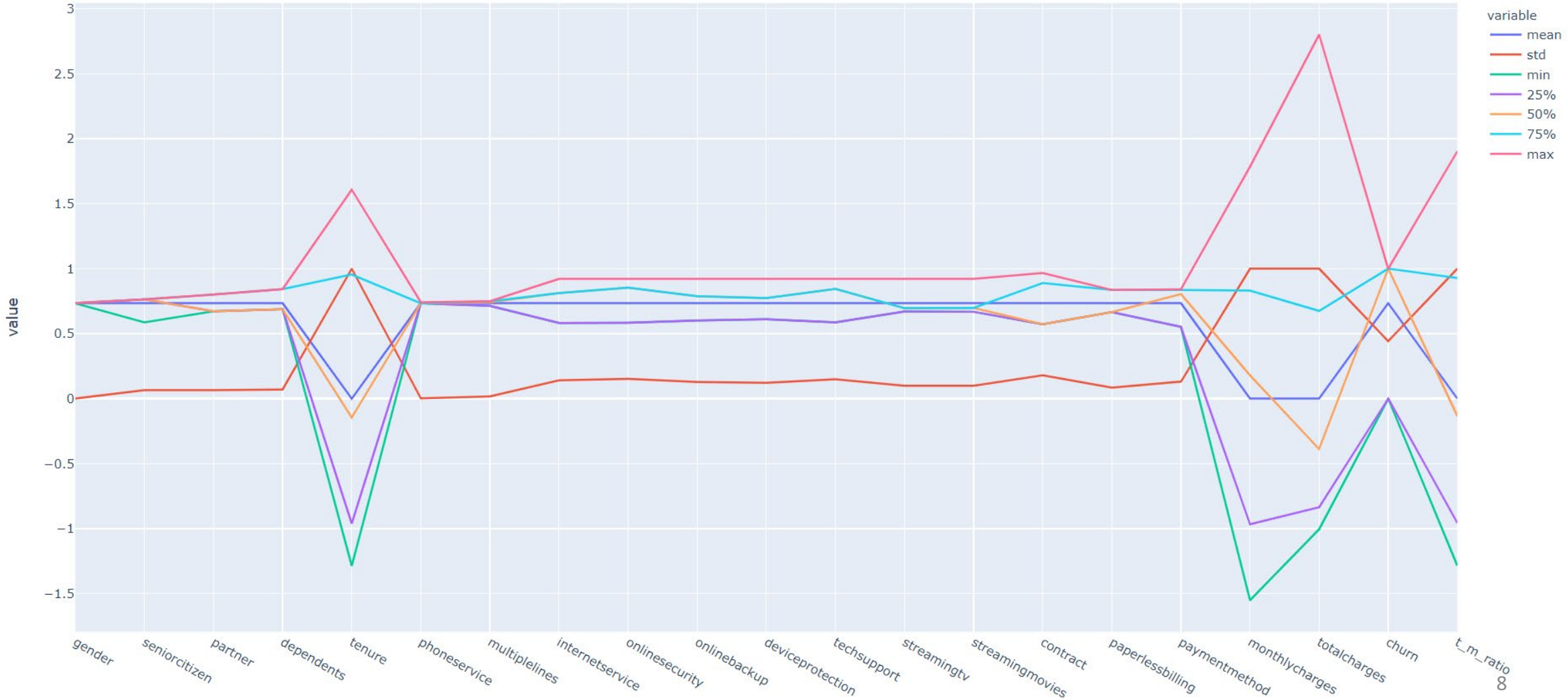




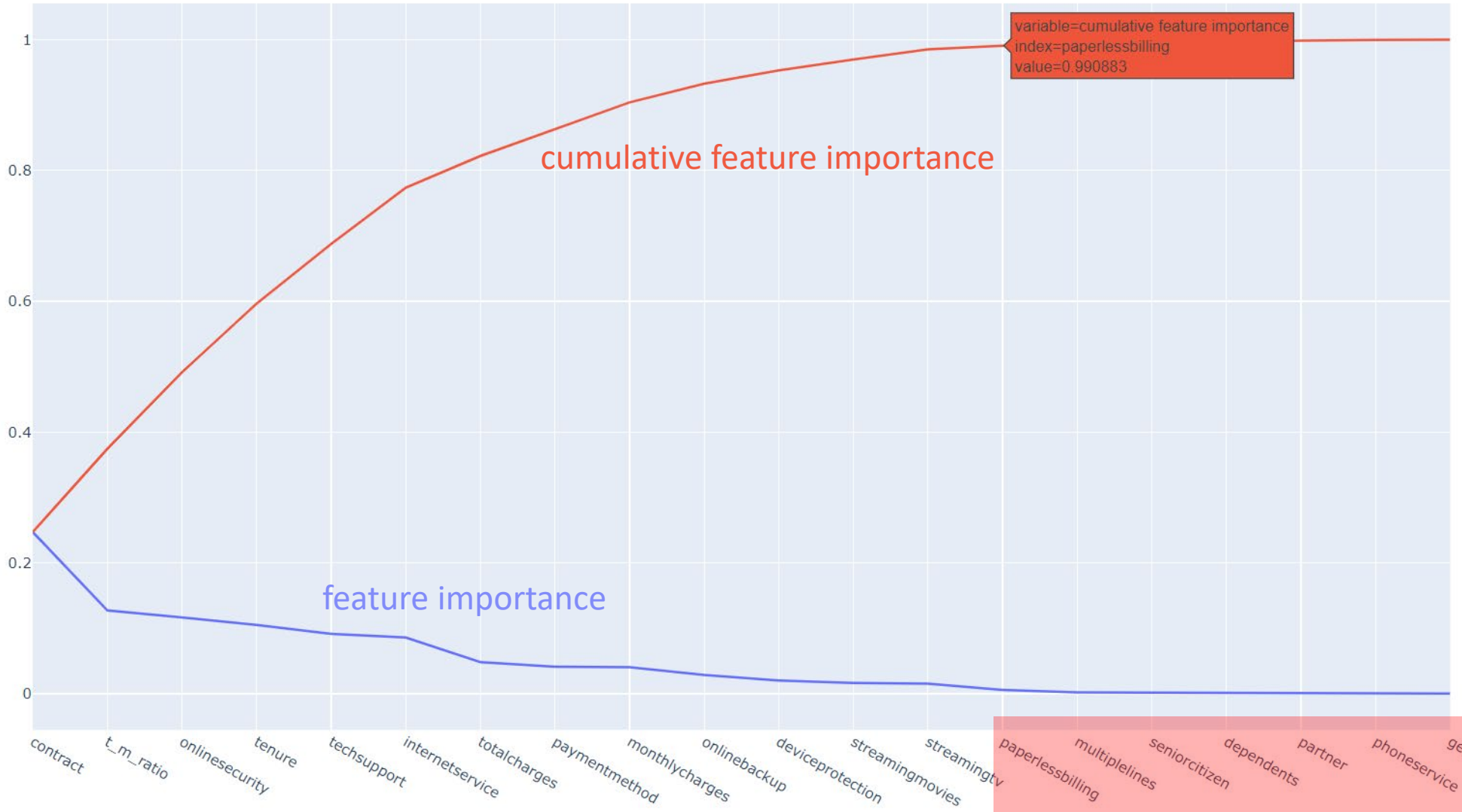


An extra feature, “t_m_ratio” is added based on the ratio of “totalcharges” over “monthlycharges” to capture the deviation between the actual payment and the scheduled payment

All features are then Z-normalized and their values are on the same order of magnitude



A quick Random Forest model is trained to remove trivial features, keeping 99% of the total feature importance



removed

It is assumed, for now, that there are no compliance considerations!

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. Preprocessing
3. **Model Training**
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. Next Steps

Three types of models are trained using an unified scikit-learn style API

For each type of model (Logistic regression, XGBoost and Random Forest)

- 5-fold cross-validation (CV) grid search is performed on the training data (80% out of the total) to identify an optimal hyper-parameter combination.
  - Note that for each CV model, 64% and 16% of the total data are used for train and validation, respectively
- A “best” model, as measured by AUC*, is then trained on the training data (80% out of the total)
- Probability of churn is then predicted using the “best” models for performance evaluation

*Area under curve (AUC) is selected due to its independence of decision cut-off

Dense Neural Network (DNN) is briefly explored but the performance is sub-optimal

- Tensorflow Keras based dense network
- Only one hidden layer
- Different learning rate, number of hidden layer units and activation function are tried
- GINI is not optimal (GINI <= 60%). It is speculated that there is not enough data as NN requires more data to be effective

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. Preprocessing
3. Model Training
4. Results
    - **A. Model Predictions**
    - B. Churn Driving Factors
    - C. Cost Benefit Analysis
5. Next Steps

Model predictions can be used to identify the "risky" customers for intervention

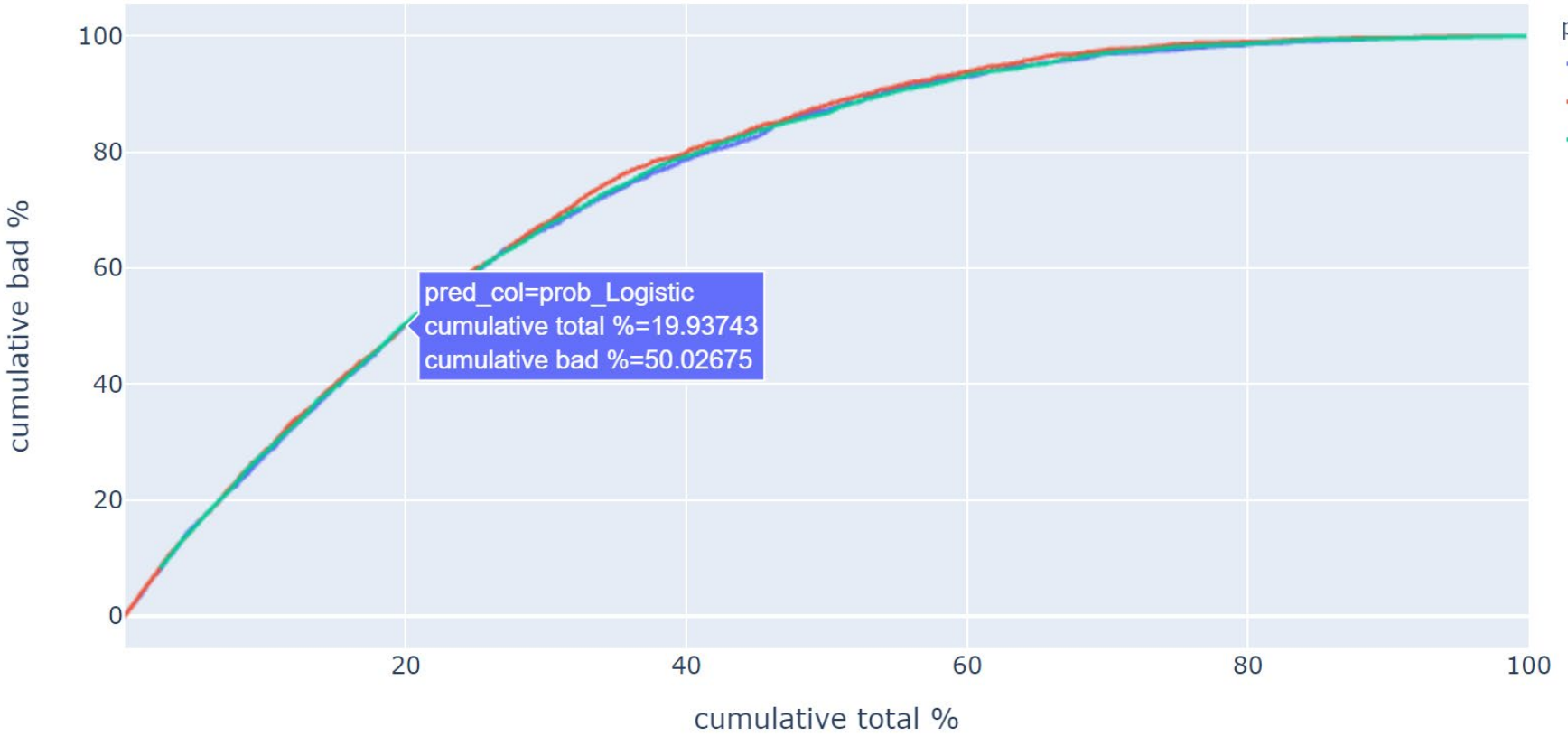
- GINI* is selected as the 1st metrics to evaluate the quality of predictions as it is independent of the decision boundary (cut-off)

model	train	test
Logistic	67.7%	71.2%
RF	70.1%	70.9%
XGB	68.4%	70.6%

- Test GINI is higher than the train GINI, suggesting potential under-fit and further tuning is required
- Performance of the three types of model is similar

**GINI = 2 * AUC – 1. It is a number between 0 and 1. A GINI of 1 corresponds to a perfect model and a GINI of 0 suggests the model provides no predicting power.*

Sensitivity is selected as the 2nd metrics to evaluate model performance*



- Similar performance of the three types of models
- 50% of churn are captured within the top 20% “risky” customers

*All data, i.e. train + test, is used instead of the test data because: 1) No over-fit on the training data; 2) There are no ample records, especially churns, in the test sample

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. Preprocessing
3. Model Training
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. Next Steps

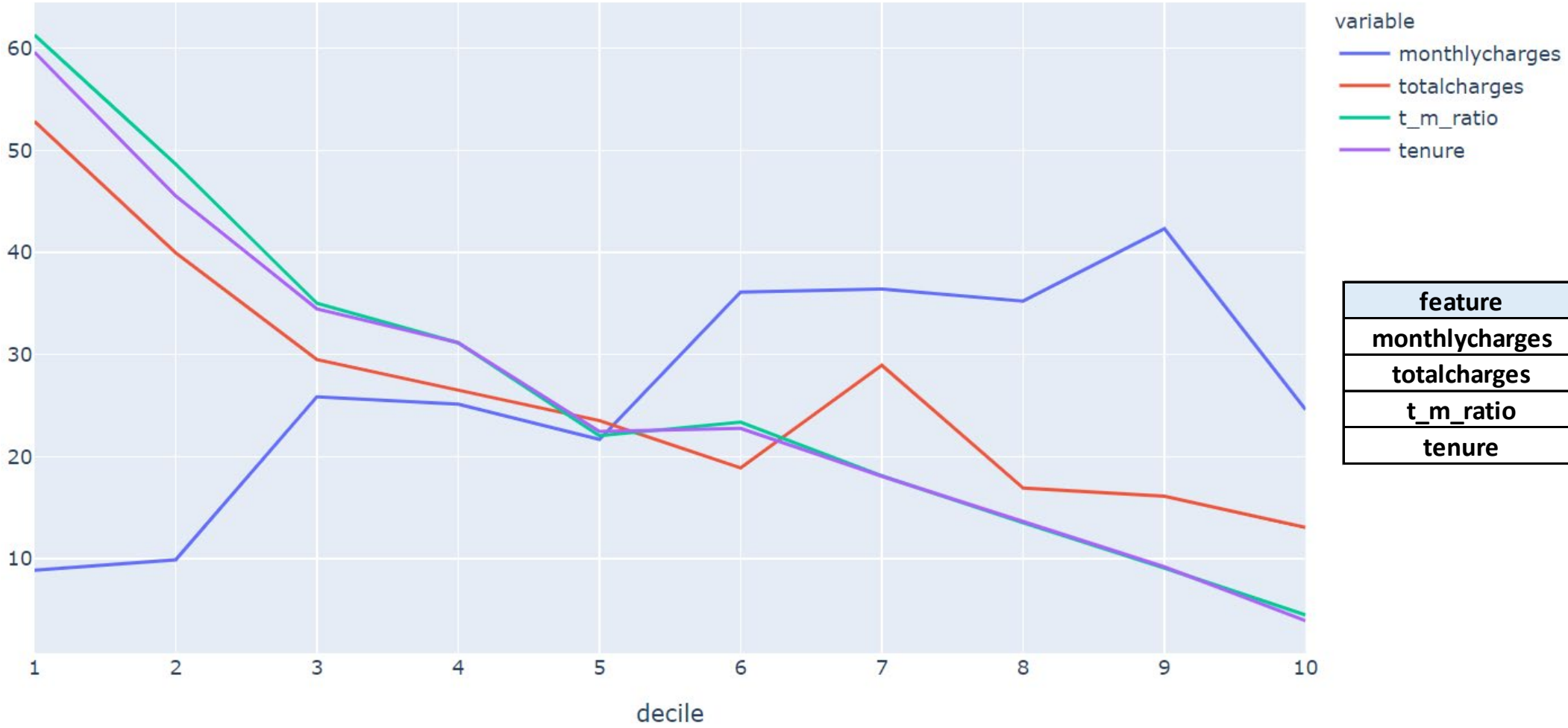
Churn rate per group per feature, together with feature importance, could be used to identify driving factors (part 1)

feature	feature importance	value	churn rate (%)
contract	32.3%	Month-to-month	42.7
contract	32.3%	One year	11.1
contract	32.3%	Two year	3.3
onlinesecurity	12.9%	No	41.7
onlinesecurity	12.9%	Yes	14.5
onlinesecurity	12.9%	No internet service	7.8
techsupport	9.4%	No	41.4
techsupport	9.4%	Yes	15.5
techsupport	9.4%	No internet service	7.8
internetservice	9.0%	Fiber optic	41.7
internetservice	9.0%	DSL	18.8
internetservice	9.0%	No	7.8
paymentmethod	2.3%	Electronic check	44.8
paymentmethod	2.3%	Mailed check	19.6
paymentmethod	2.3%	Bank transfer (automatic)	16.4
paymentmethod	2.3%	Credit card (automatic)	15.8
onlinebackup	1.5%	No	39.9
onlinebackup	1.5%	Yes	21.2
onlinebackup	1.5%	No internet service	7.8
streamingmovies	0.9%	No	33.2
streamingmovies	0.9%	Yes	30.2
streamingmovies	0.9%	No internet service	7.8
deviceprotection	0.9%	No	38.9
deviceprotection	0.9%	Yes	22.6
deviceprotection	0.9%	No internet service	7.8
streamingtv	0.8%	No	33
streamingtv	0.8%	Yes	30.4
streamingtv	0.8%	No internet service	7.8

Churn rates for categorical features

Churn rate per group per feature, together with feature importance, could be used to identify driving factors (part 2)

Churn rates for numerical features by deciles (%)



feature	feature importance
monthlycharges	4.7%
totalcharges	4.0%
t_m_ratio	12.7%
tenure	9.1%

Actions to improve customer stickiness/loyalty **BEFORE** they become "risky", are feasible for some driving factors but not for others

Note the suggestions here are nothing but **educated guesses** given the limited information available on the business's operation

Yes

- Tech Support
- Monthly Charges, Internet Service
- Payment Method

No

- Contract
- Total Charges, Tenure

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. Preprocessing
3. Model Training
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. Next Steps

The benefit of introducing a model and its cut-off, can be estimated with some key cost/benefit/likelihood information provided

- Cost/Benefit associated with confusion matrix components
  - TP: Intervene “risky” customers --- what is the CLV increase?
  - TN: Ignore loyal customers
  - FN: Ignore “risky” customers --- what is the CLV loss?
  - FP: Intervene loyal customers
- Success rate, cost and effect of intervention
- Other cost/benefit to consider
  - Development, deployment, monitoring and update of model and services in future
  - Integration cost of related departments
  - Business branding/reputation

Contents

1. Exploratory Analysis (EPA) and Problem Statement
2. Preprocessing
3. Model Training
4. Results
    A. Model Predictions
    B. Churn Driving Factors
    C. Cost Benefit Analysis
5. **Next Steps**

Possible actions towards a better predictive model & service

- Obtain more data in terms of volume and ideally variety
- More extensive feature engineering
- Further tuning of models
- Clarify the operation of related business departments
- Quantify the benefit of the model & service
  - First, retrospectively
  - Then A/B test
- Create API/Service for integration into existing services