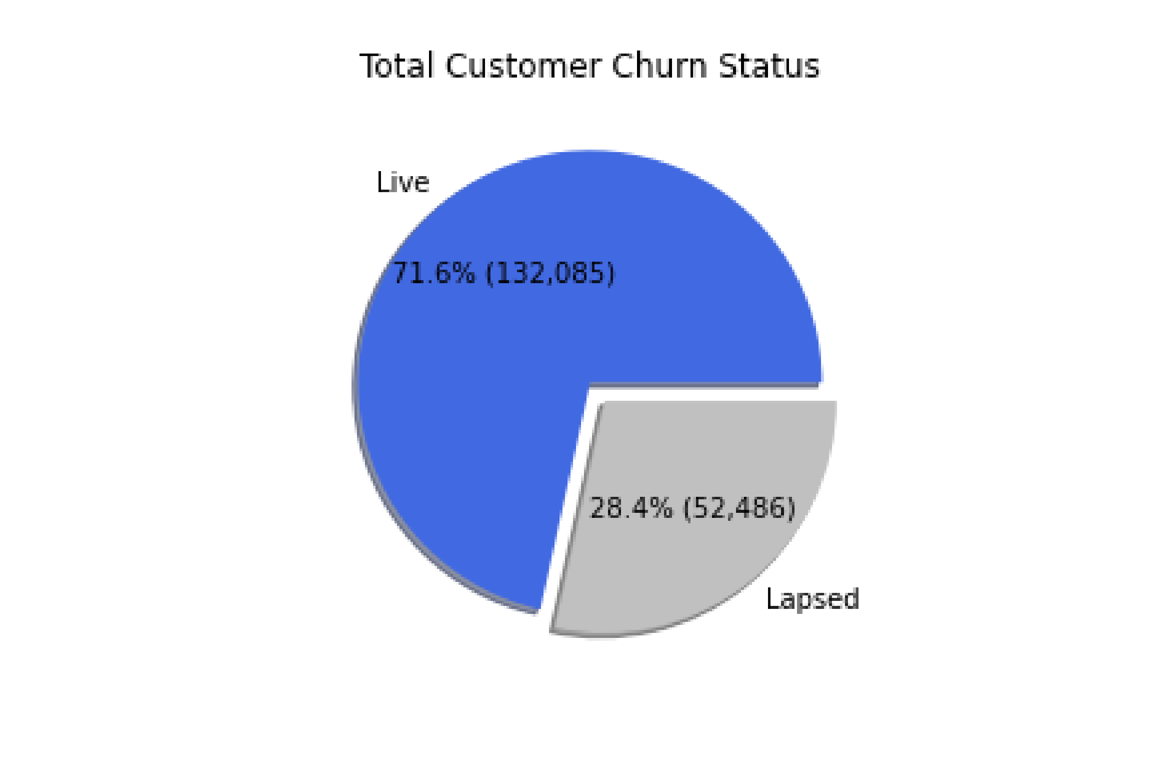
Home insurance analysis

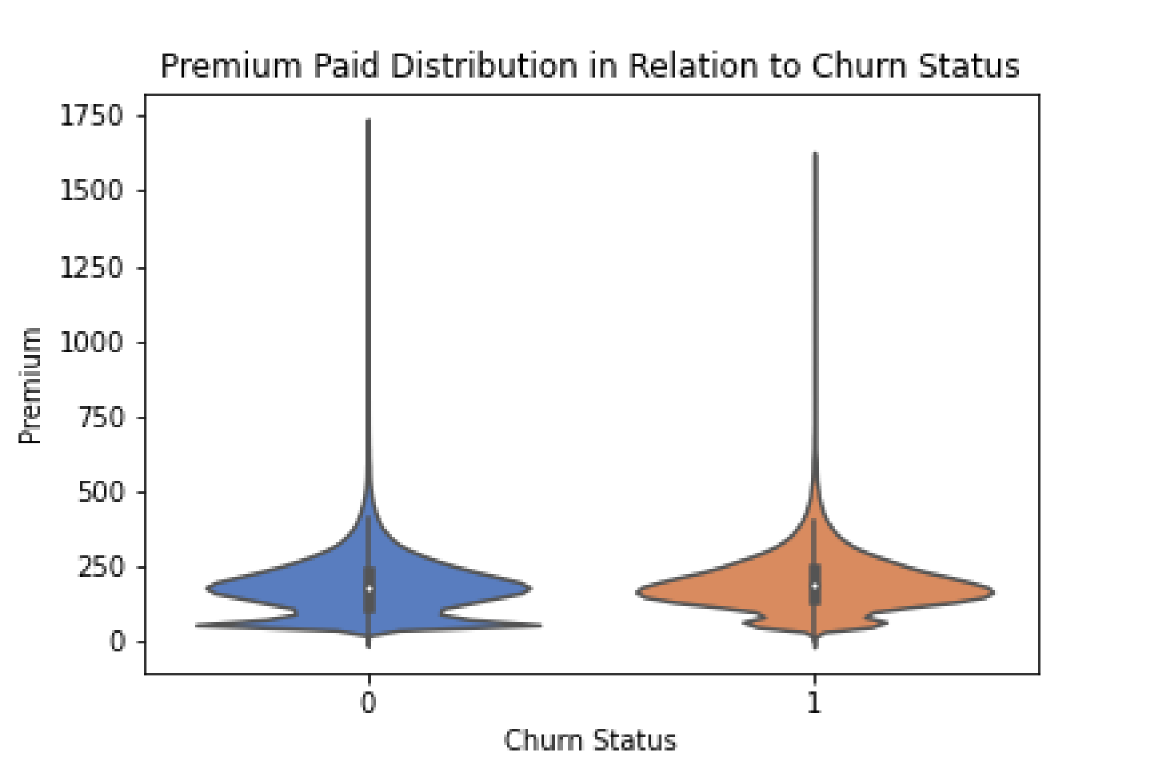
datasource: https://www.kaggle.com/ycanario/home-insurance

Daniel Balseanu

Data Analysis

Policy Status



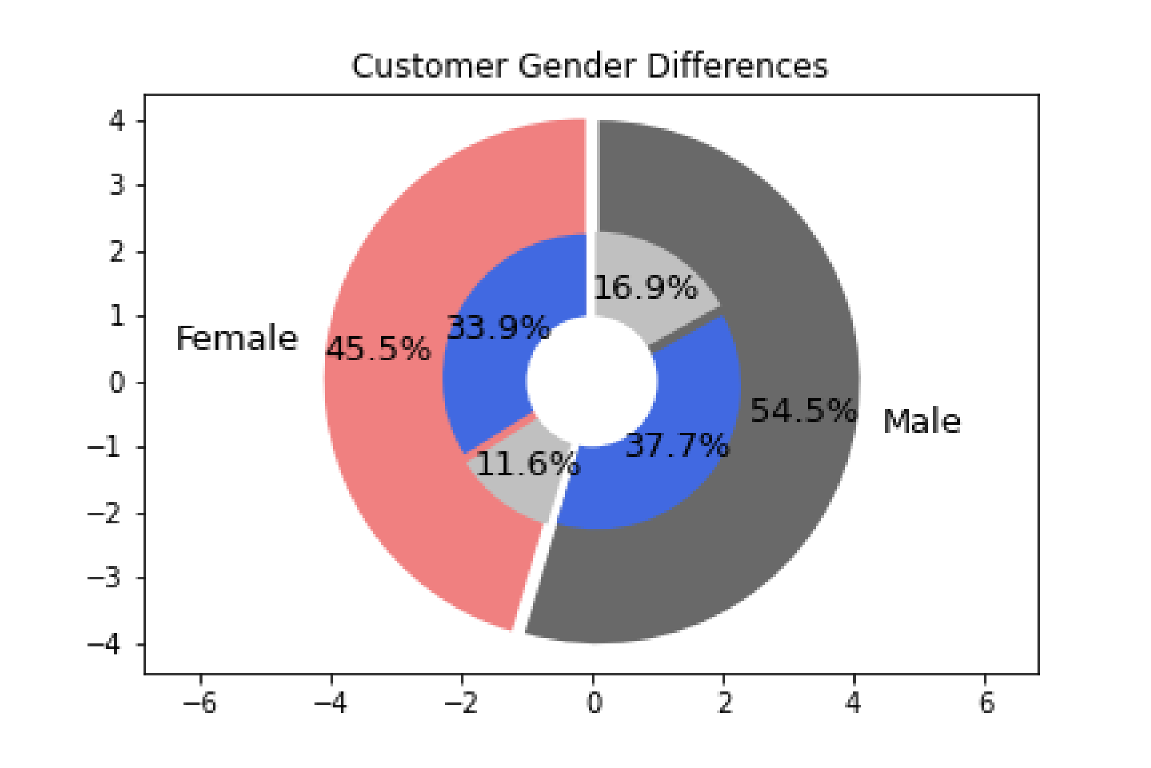


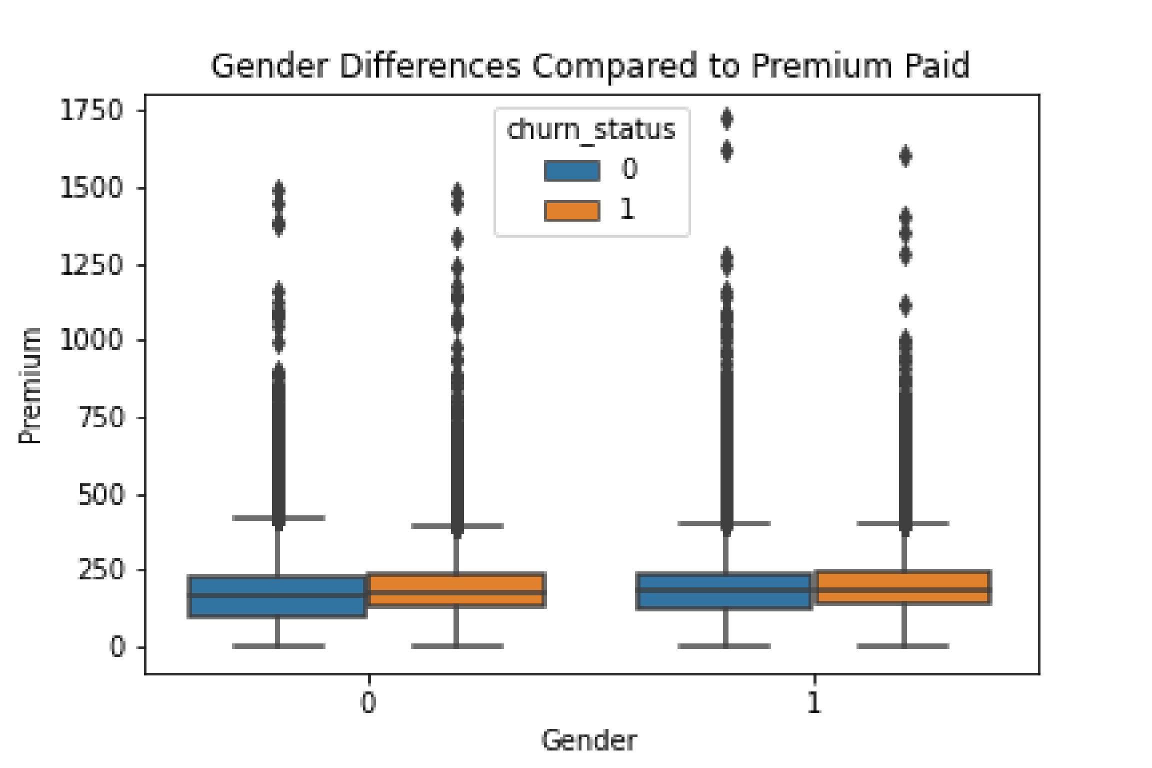
Initial Observations

- Total of 184,571 Policies analysed after data cleansing, 28.4% total churned.
- Premium would be an obvious culprit: distribution shows skewed number of active customers with a small premium.
- No significant statistical correlations between Policy Status and other features.

Data Analysis

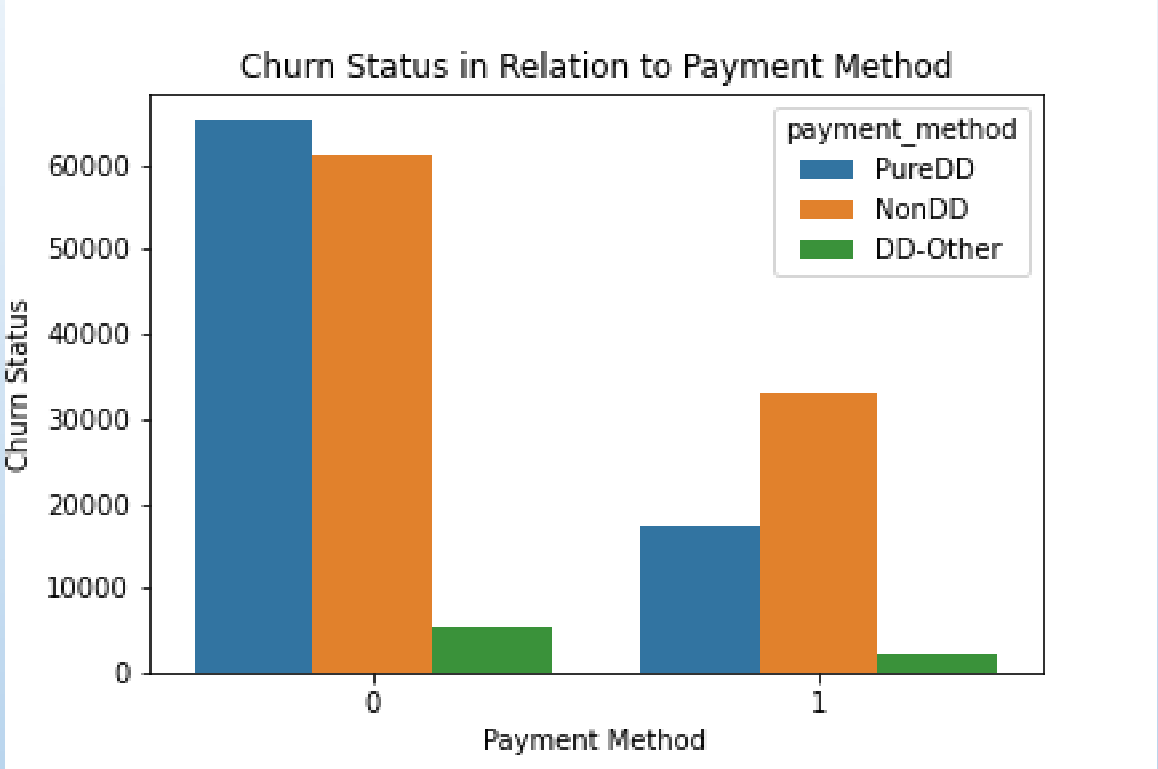
Gender





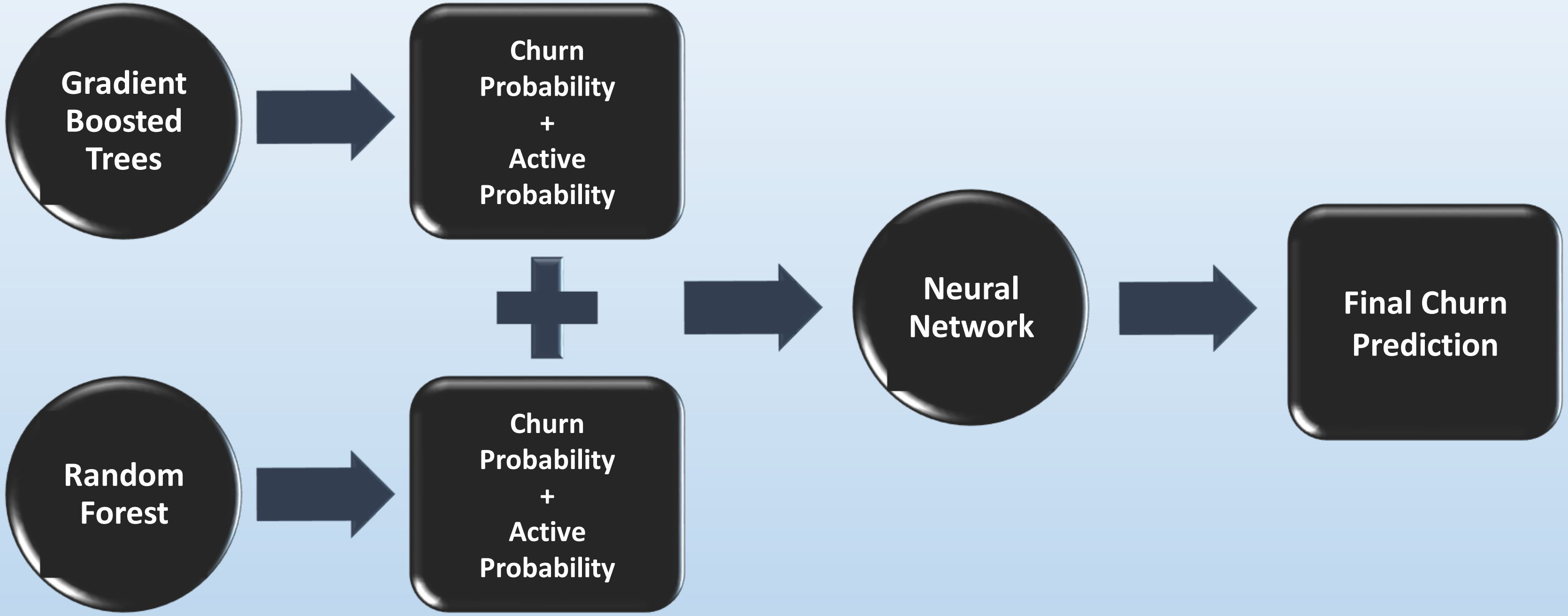
- Is gender relevant? E.g. Risk related tendencies
- A proportionally greater part of the males are churned when compared to females.
- In relation to premium, tendency for active policy female customers to pay less premium.

Data Analysis
Payment Method



- Recursive payment methods generally reduce friction and potential churn.
- Greater proportion of non direct debit customers are no longer active.

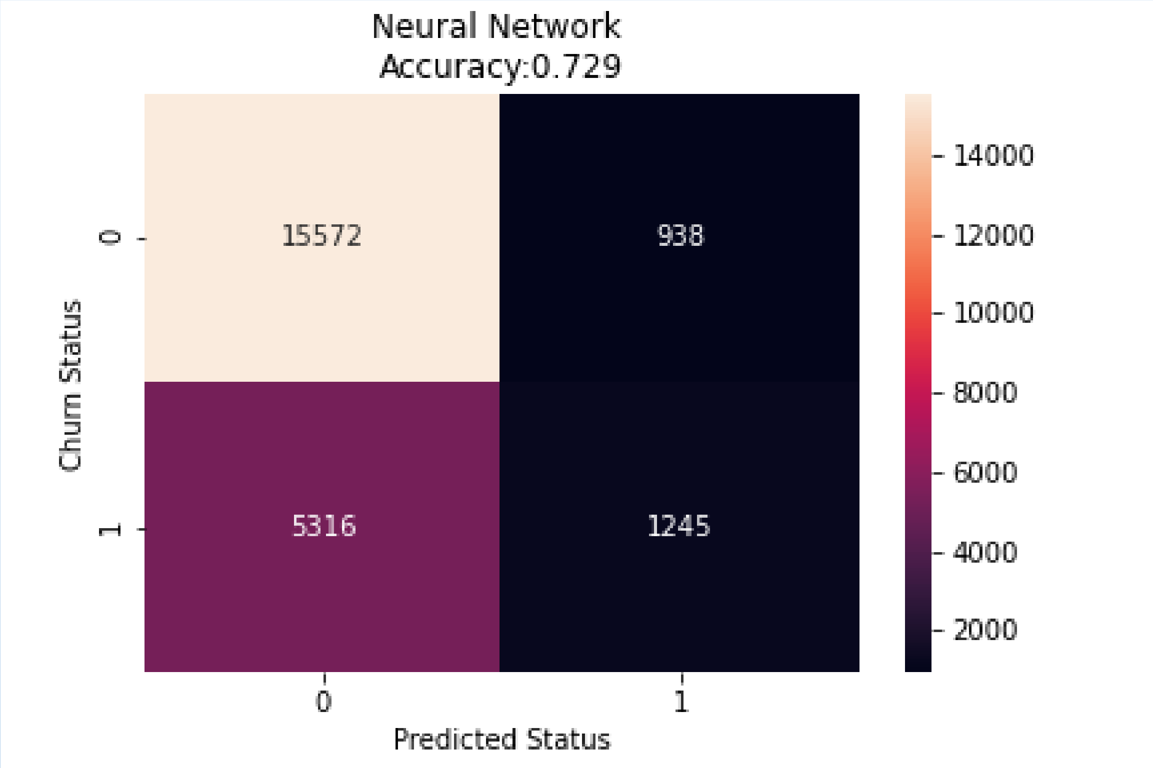
Model Architecture



- 1^{st} Predictive Layer: XGBoost and Random Forest models provide initial class probabilities.
- 2^{nd} Predictive Layer: Stacked Neural Network aggregates initial predictions.
- Data split into multiple parts: train (75%), validation (12.5%), test (12.5%).
- Data preparation: mean encoding categorical variables, binary encoding variables with 2 levels, minimal outlier removal and missing data imputation for Random Forest model.

Model Results

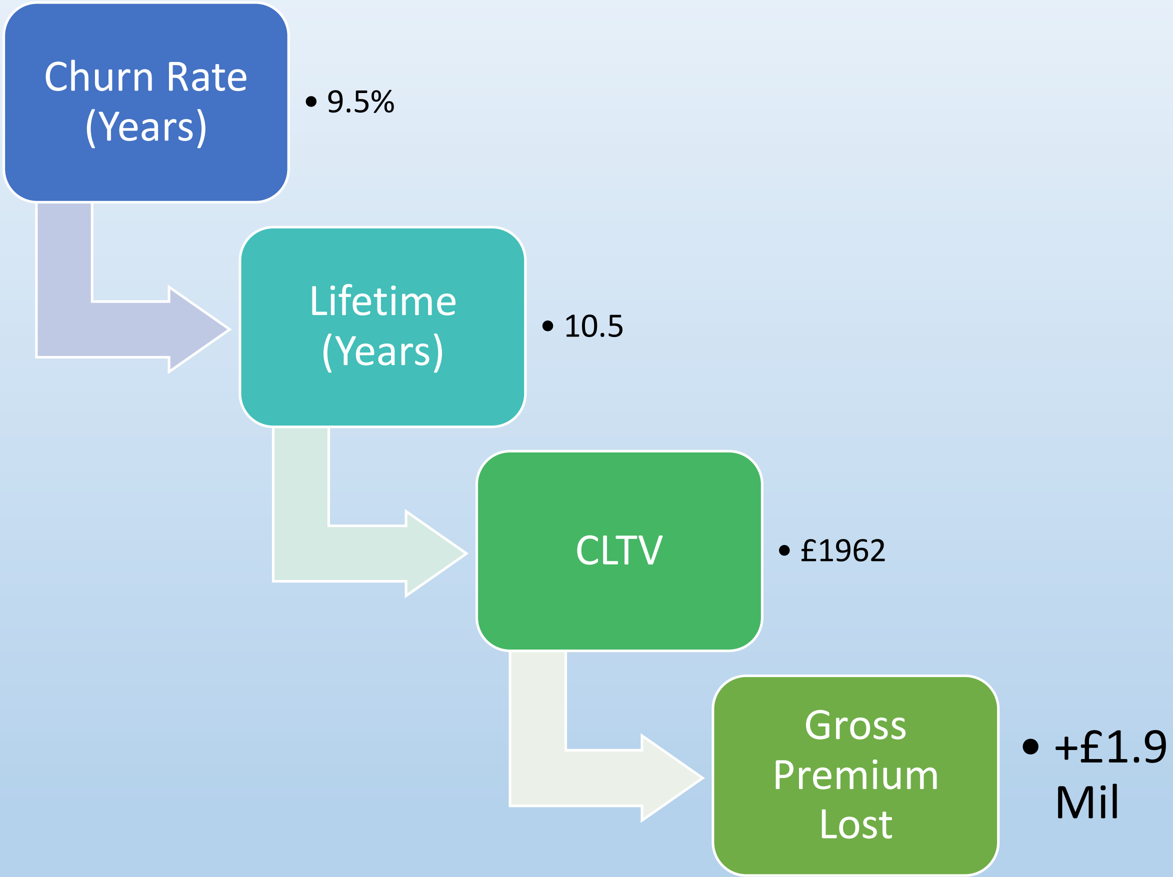
Test data



Metric	Random Forest	XGBoost	Neural Network
Recall	0.02	0.16	0.19
Precision	0.8	0.58	0.57
F1 Score	0.05	0.25	0.28

- Recall: How many actually churned customers did we predict?
- Precision: From the customers that we predicted will churn, how many actually churned?
- F1 Score: Blends Recall and Precision in a general score
- On the test set models perform only slightly worse, F1 score on the training set being 0.30

Cost Benefit Analysis



- Given the total Live vs Lapsed customers, the estimated Churn rate is 9.5% with a life time of 10.5 years.
- Based on the test set predicted churns, gross premium value lost is £1.9 mil.
- Potential estimated net value 10%: £190,243 or £153 per customer retained.

Further considerations:

- Recursive challenge: if we decrease churn rate, then the LTV calculations are under estimated while also having fewer lapsed policies to predict.
- Is there a cost to customers that are incorrectly predicted as churned?

Conclusion

Are the models created on an inherently biased premise?

- The data provided is collected post customer lapse – does it actually have predictive power?
- Need to establish causation – what happened prior to churn?

Different Perspective

Customer State 1 

- Married
- No Claims
- Good Financial Situation
- Policy year 1

→

Customer State 2 

- Divorced
- Claims made
- Mediocre Financial Situation
- Policy year 2

→

Customer Lapsed

Customers have a specific timeline leading to churn. Modelling data this way allows potential for different approaches:

- Generative Bayesian Multi Armed Bandit
- Reinforcement Learning

Other potential features:

- Interactions with the insurance provider: complaints, emails, phone calls, claims, website visits
- Socio-economic dimensions: income, family size, moving home

Other Notes

- Code modular, documentation provided.
- All models and reports are saved, results can be reproduced.
- Plug in new data.csv and run main.py it will produce a new up to date model.
- XGBoost model hyper parameter tuning done with cross validation implementation.
- What about the unit tests?
- Feature selection can be improved, difficult to optimize due to few continuous dimensions.
- Predictions thresholds at 0.5, however can be altered to potentially increase Recall at the cost of Precision (PR Curve more useful than ROC for this analysis).