Porto Seguro’s Safe Driver Prediction

Team 3

Euan Enticott 160002105

Supanuch Juengsanguansit 180018506

Xuan Zhang 180026889

Chun Paul Ho 180029517

Sam Reilly 180016064

Key Discussion Area

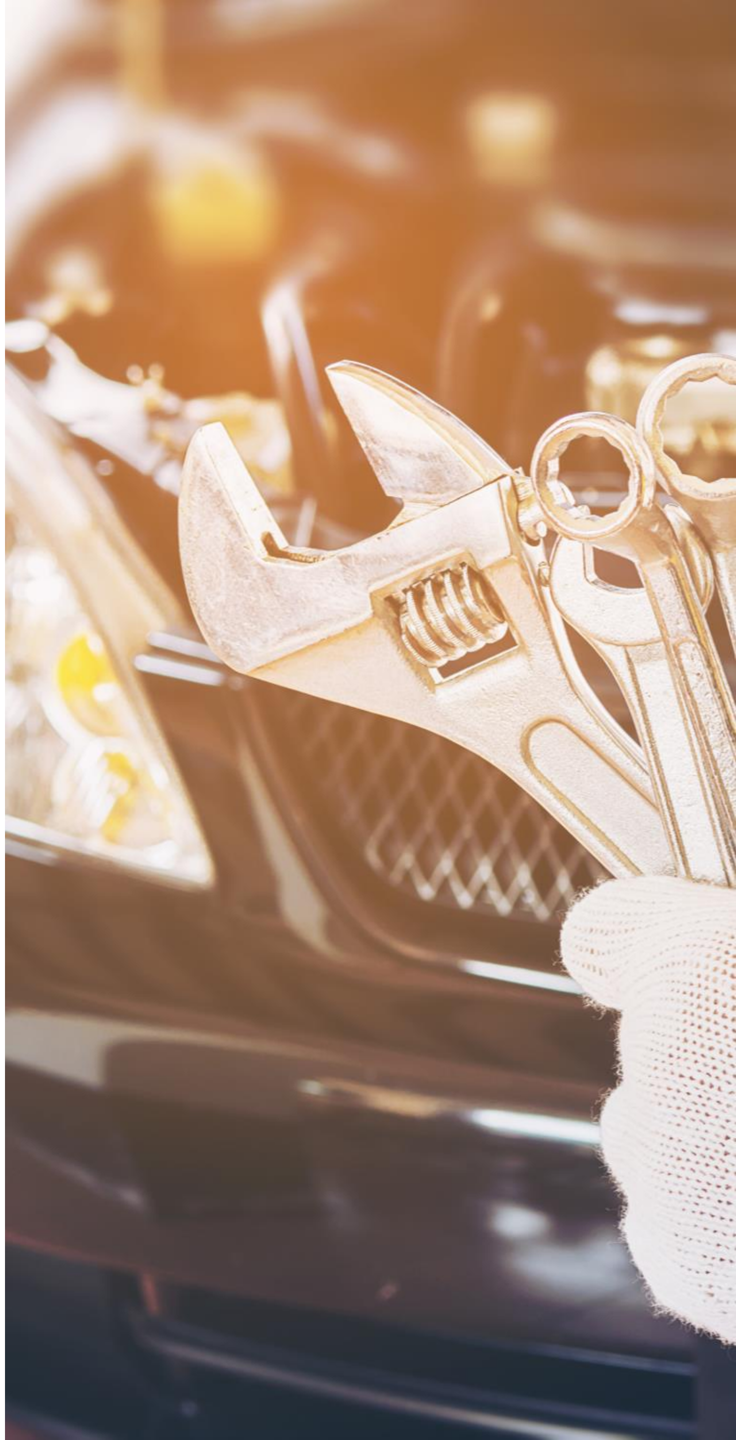
- Overview of the Problem
- Initial Data Analysis Exploration
- Outline Modeling Approach
- Evaluation
- Conclusion and Discussion

Overview of the Problem

❑ Identify the individuals who have a high risk of making an insurance claim within the next year

❑ Beneficial results would include:

- Saving based on offering “high risk” individuals higher premiums or refusing to offer insurance package
- Target “low risk” customers with competitive rates



Data Analysis Exploration

- Oversaturation of customers who do not make a claim
- Highlight the extent of the challenge, as an example, a model that predicts the correct result 97% of the time

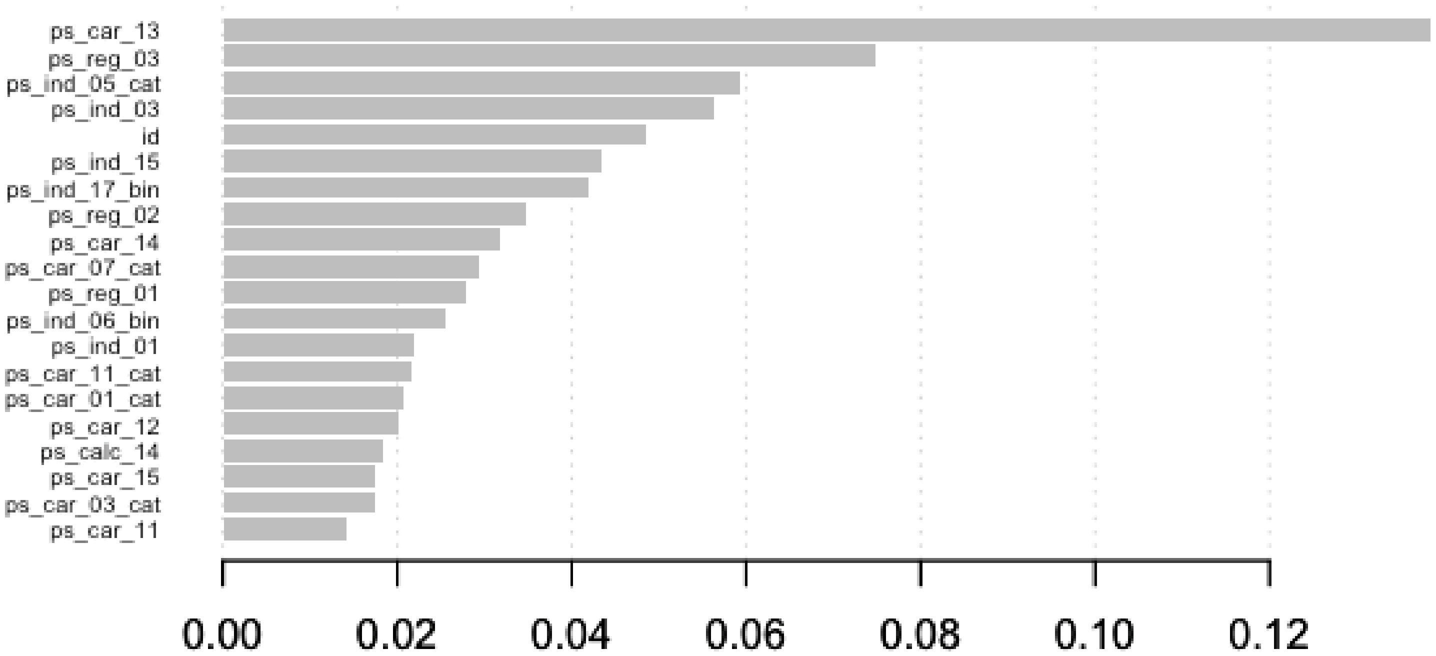
Data Analysis Exploration

- Oversaturation of customers who do not make a claim
- Highlight the extent of the challenge, as an example, a model that predicts the correct result 97% of the time

However, all the model has to do is predict every individual to make a claim, which in the real world generates no value.

Data Analysis Exploration

- Delete all “calc” columns



- Both integer & one-hot encoding

- Generate a feature to identify if missing values were present or not

Our Modeling Approach

- Consider a wide range of candidate models
- Hyper-tune parameters of best individual model
- Generate an Ensemble model & validate using cross validation
- Translate Model predictions into a solution which gives Porto Seguro a competitive edge in a growing market

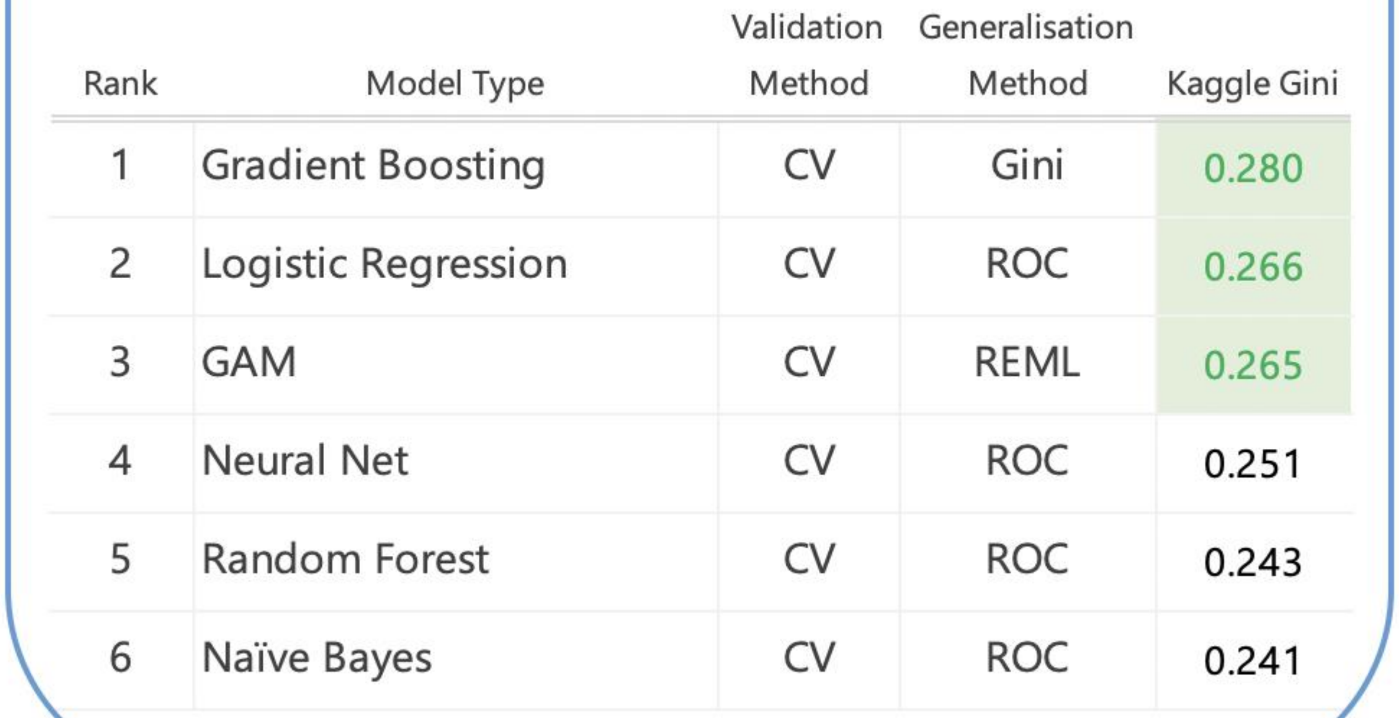
Candidate Models

- 6 models have been tested
  - Gradient boosting
  - GAM
  - Logistic regression
  - Neural Network
  - Random forest
  - Naïve Bayes
- CV for validation method
- Normalized Gini for generalization performance

Performance of Candidate Models

Step 1: Consider wide range of Candidate Models

Generate set of models and rank by Gini score verified by upload to Kaggle on Test data.



Rank	Model Type	Validation Method	Generalisation Method	Kaggle Gini
1	Gradient Boosting	CV	Gini	0.280
2	Logistic Regression	CV	ROC	0.266
3	GAM	CV	REML	0.265
4	Neural Net	CV	ROC	0.251
5	Random Forest	CV	ROC	0.243
6	Naïve Bayes	CV	ROC	0.241

Hyper-Tune

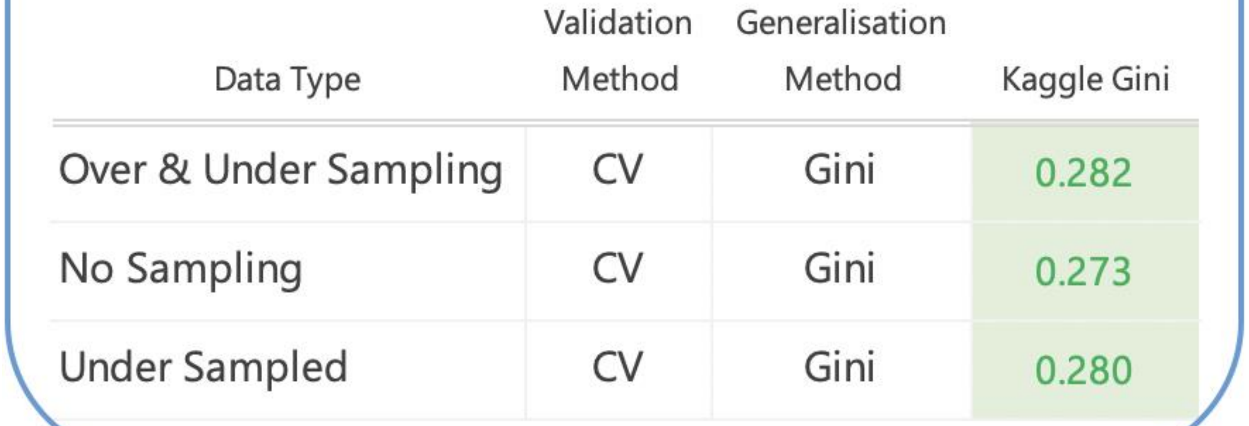
User Input

@parameter_TuneGrid

Example for Gradient Boosting:
- ETA: range from 0.02 to 0.2
- Sub Sample: range from 0.4 to 0.8

Step 2: Hyper-tune model parameters

Using best individual model, hyper-tune parameters using user defined @parameter_TuneGrid.



Data Type	Validation Method	Generalisation Method	Kaggle Gini
Over & Under Sampling	CV	Gini	0.282
No Sampling	CV	Gini	0.273
Under Sampled	CV	Gini	0.280

Ensemble Model

User Input

@GiniThreshold = 0.26

Step 3: Generate Ensemble Model

Every model that achieves Gini score on Kaggle above @GiniThreshold variable should be included.

Ensemble model contains:

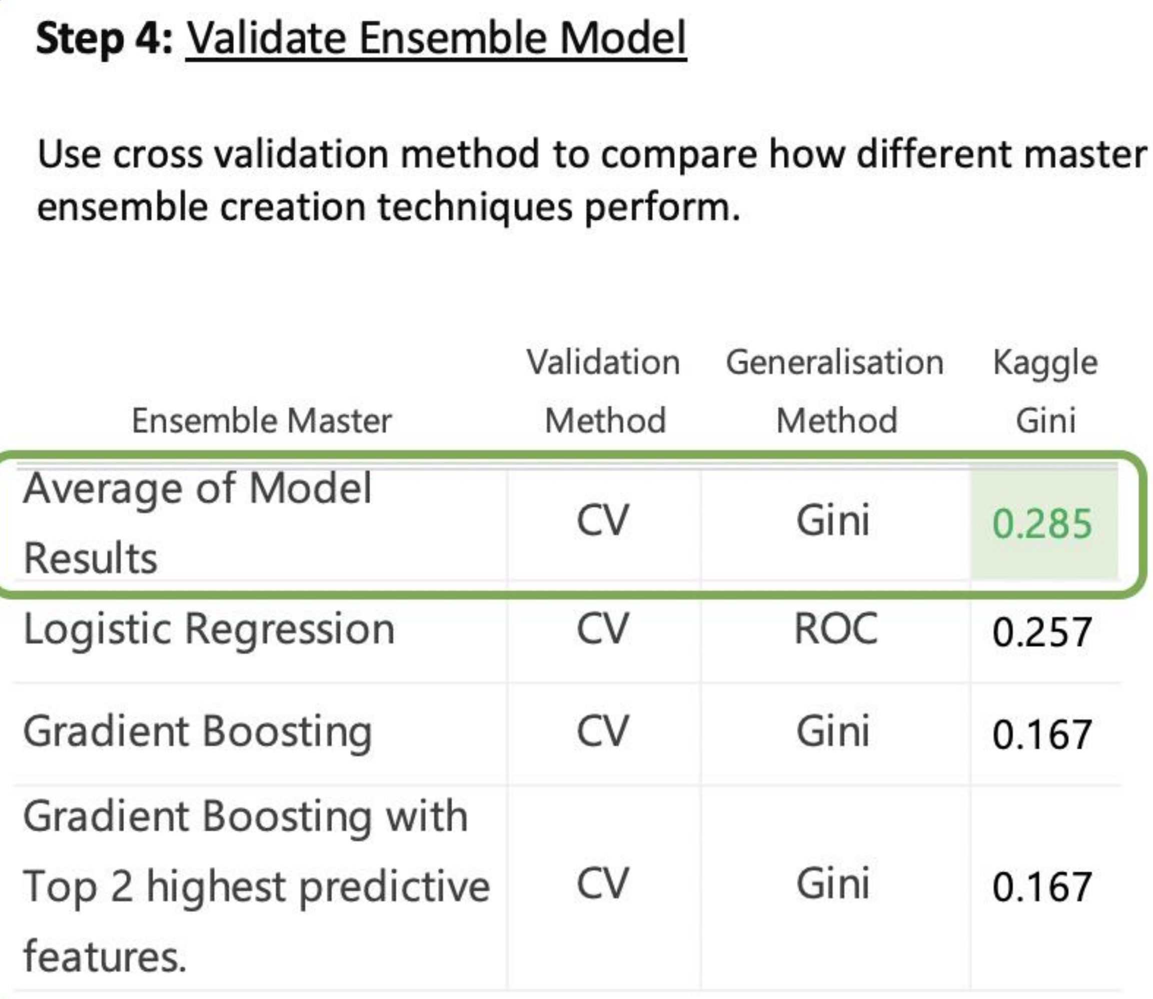
i) 3x Gradient Boosting

ii) Logistic Regression

iii) GAM

Ensemble Model

Validation

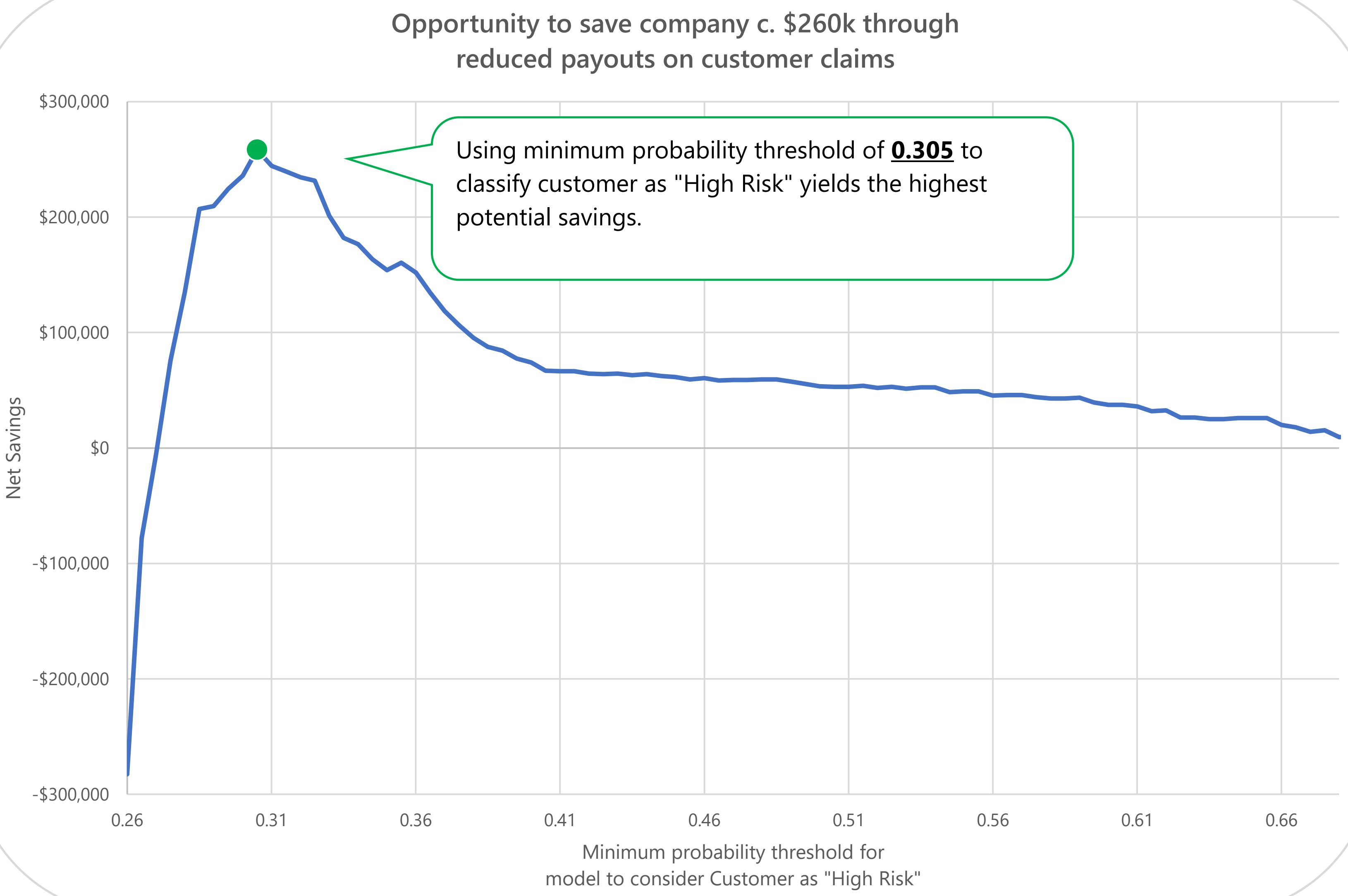


Step 4: Validate Ensemble Model

Use cross validation method to compare how different master ensemble creation techniques perform.

	Validation	Generalisation	Kaggle
Ensemble Master	Method	Method	Gini
Average of Model Results	CV	Gini	0.285
Logistic Regression	CV	ROC	0.257
Gradient Boosting	CV	Gini	0.167
Gradient Boosting with Top 2 highest predictive features.	CV	Gini	0.167

Potential Financial Savings

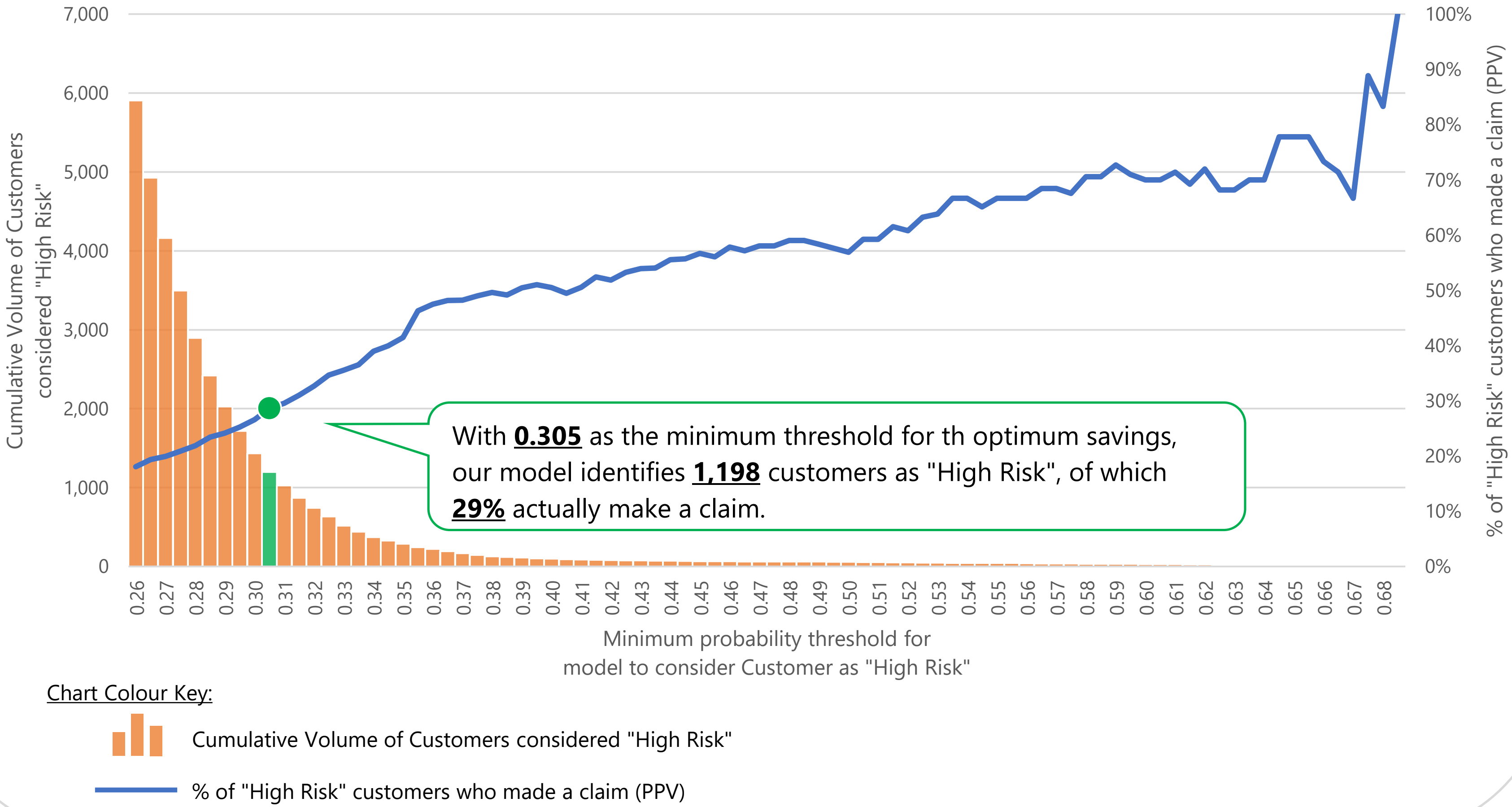


Opportunity to save company c. $260k through reduced payouts on customer claims

Using minimum probability threshold of **<u>0.305</u>** to classify customer as "High Risk" yields the highest potential savings.

Potential Financial Savings

Distirbution of cumulative customer volume & **PPV** as minimum probability threshold for model to consider customer as "High Risk" increases



With **0.305** as the minimum threshold for th optimum savings, our model identifies **1,198** customers as "High Risk", of which **29%** actually make a claim.

Chart Colour Key:

- Cumulative Volume of Customers considered "High Risk"
- % of "High Risk" customers who made a claim (PPV)

Conclusion and Discussion

- Created an Ensemble model using 6 individual models
- Scaling Financial saving to client base of 1.4 million yields potential annual savings of $600K
- Developed script that will automatically apply our ensemble model to new data which generates:

  - Probabilities that customers will make a claim

  - Classify* those that are “high risk”

* {based on 0.305 value discussed earlier however this is an available parameter that can be updated at anytime}

Thank you for your time.

Any questions?