



SMOKER PROPENSITY AND SMOKER LIAR MODEL

Minh Bui, Jeff Hevrin

Brad Zelnio, Matt Hartman

Introduction

Smoker Propensity Model: Predicts which applicant is a smoker

Smoker Liar Model: Predicts which applicant is a smoker liar

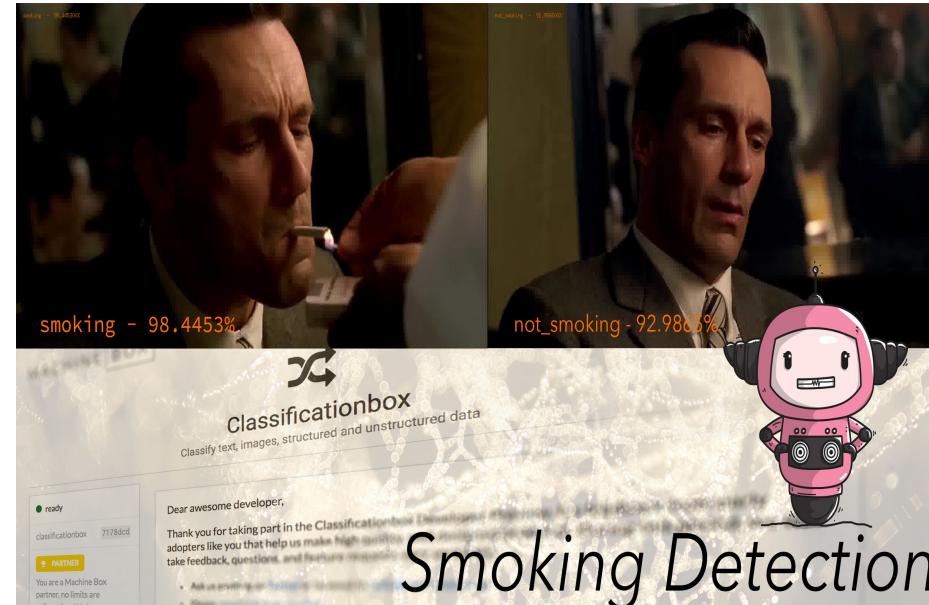
Model characteristics:

- Same performance metrics but different distribution and results
- Follow same procedure:
 - Problem definition
 - Data acquisition
 - Modeling
 - Evaluation
 - Features
 - Experimentation



PROBLEM DEFINITION

- An insurance fraud: Applicants giving false information about their tobacco use.
- Build two models to support the accelerated underwriting process in flagging applicants who are likely to be smokers and smoker liars.
- Supervised Machine Learning: figuring out the patterns based on input-output pairs and then apply the pattern to other inputs to find outputs.

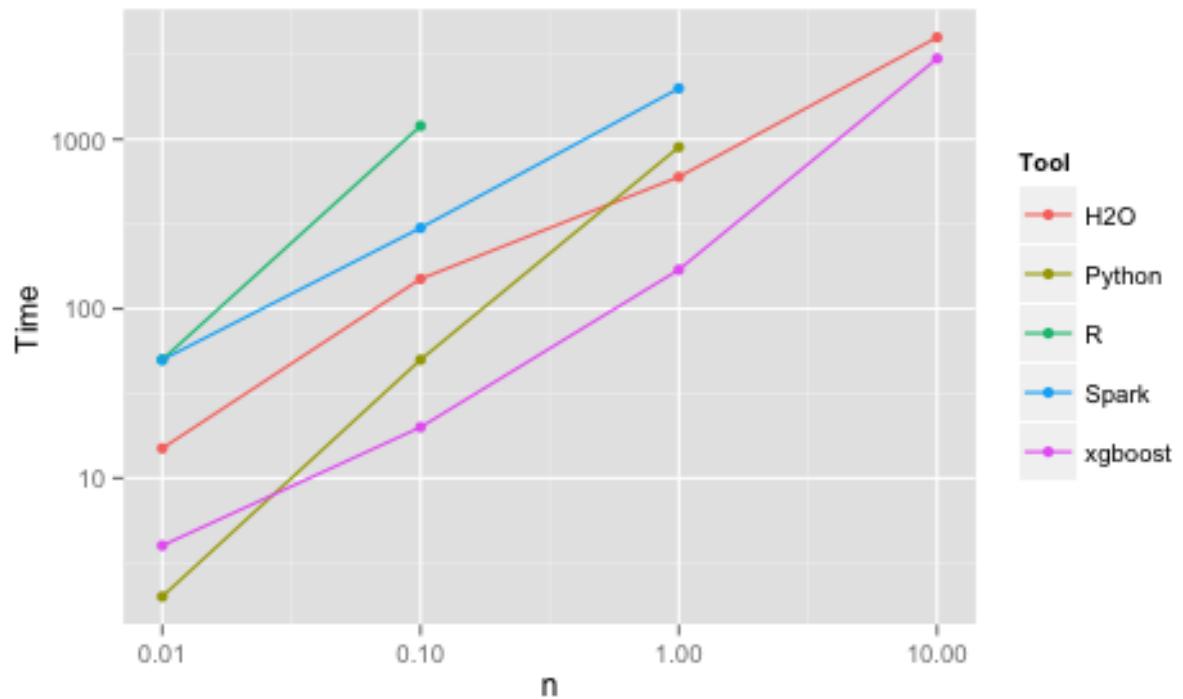


DATA ACQUISITION

- Data collected from: Application part A, application part B, lab results, therapeutics, and prescriptions
 - Gender
 - Height
 - Weight
 - Marital status
 - Age
 - Product name
 - Birth country
 - Income
 - Occupation
 - Range values & Range texts
 - Therapeutics
 - Part B application questions
- The final model has ~72000 rows and 17 columns, 7 of which contain categorical (non-numerical) values.



MODELING



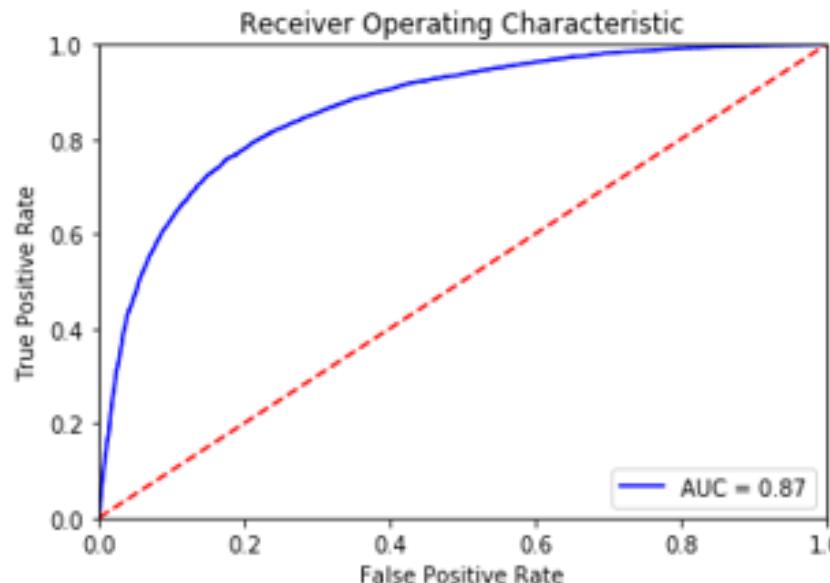
Extreme Gradient Boosting (XGBoost)

- **Fast:**
"I also tried xgboost, a popular library for boosting which is capable to build random forests as well. It is fast, memory efficient and of high accuracy" – Szilard Pafka
- **Popular:** go-to algorithm used for data science competition
"When in doubt, use xgboost." – Owen Zhang

EVALUATION

AUC Score: Area Under the Curve

- Measures how well the model distinguishes between two target variables
- Should be $0.5 < \text{AUC} < 1$, 1 is for 100% accurate prediction while 0 is 100% wrong prediction.



Example of a AUC curve

Accuracy score:

$$\text{Accuracy} = \frac{\text{Number of accurate predictions}}{\text{Total number of predictions}}$$

- Doesn't work for imbalanced datasets

	Predicted True	Predicted False
Real True	850	35
Real False	45	70

$$\frac{70 + 850}{1000} = 92\%$$

Example of accuracy score with imbalanced dataset

EVALUATION

Precision score:

$$\text{Precision} = \frac{\text{True Positive}}{\text{Total Predicted Positive}}$$

What proportion of positive identifications was correct?

Recall score:

$$\text{Recall} = \frac{\text{True Positive}}{\text{Total Actual Positive}}$$

What proportion of actual positives was identified correctly?

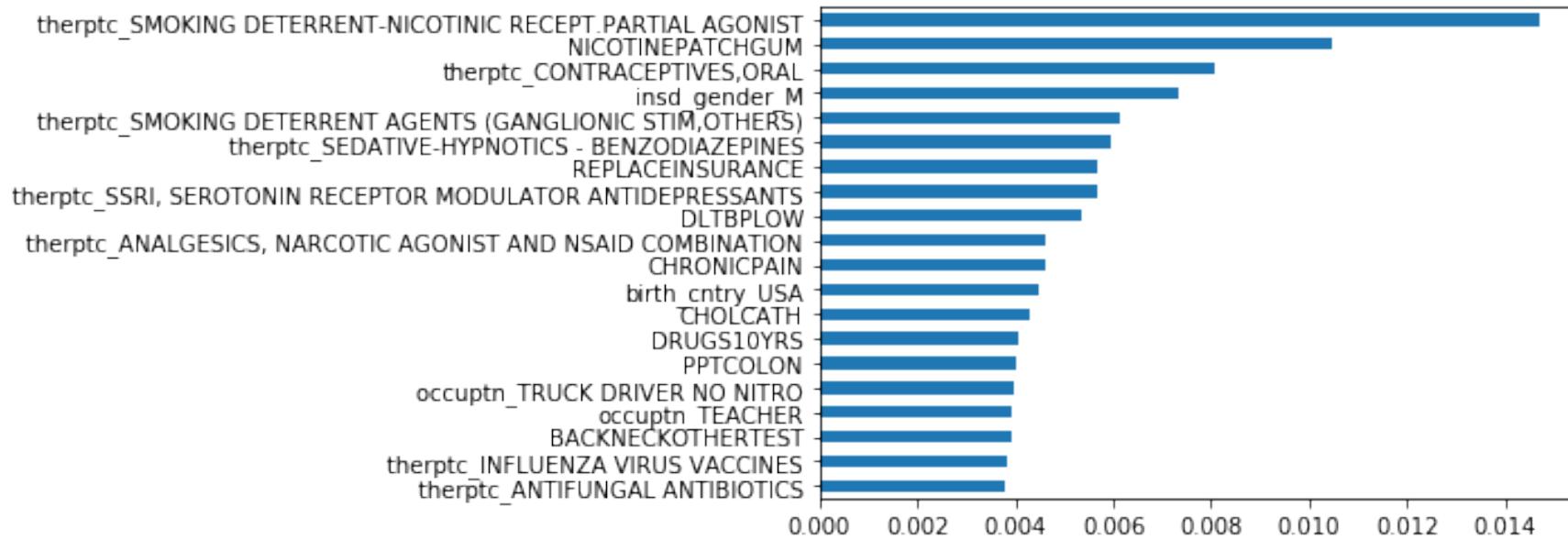
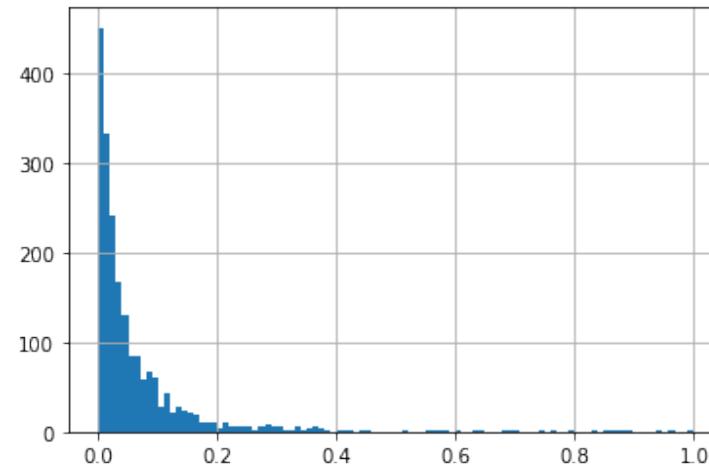
Confusion Matrix		Predicted		
		FALSE	TRUE	
Actual	FALSE	True Negative (TN)	False Positive (FP)	Precision
	TRUE	False Negative (FN)	True Positive (TP)	

Precision: The ratio of True Positives to the sum of True Positives and False Positives.

Recall: The ratio of True Positives to the sum of True Positives and False Negatives.

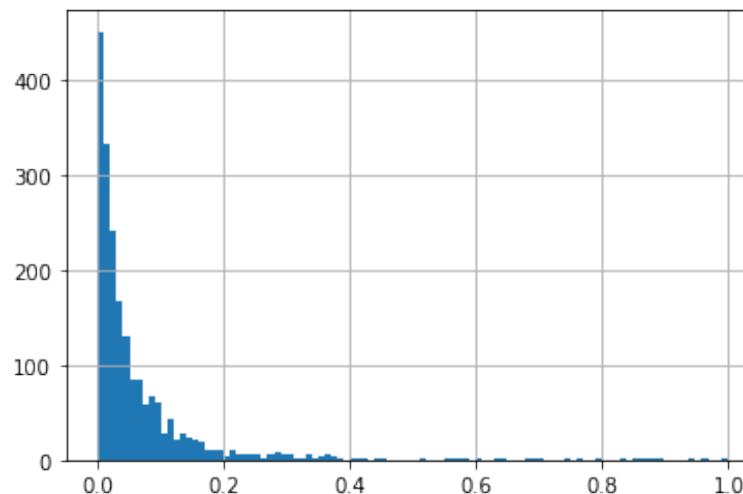
SMOKER PROPENSITY MODEL

AUC score: 0.74



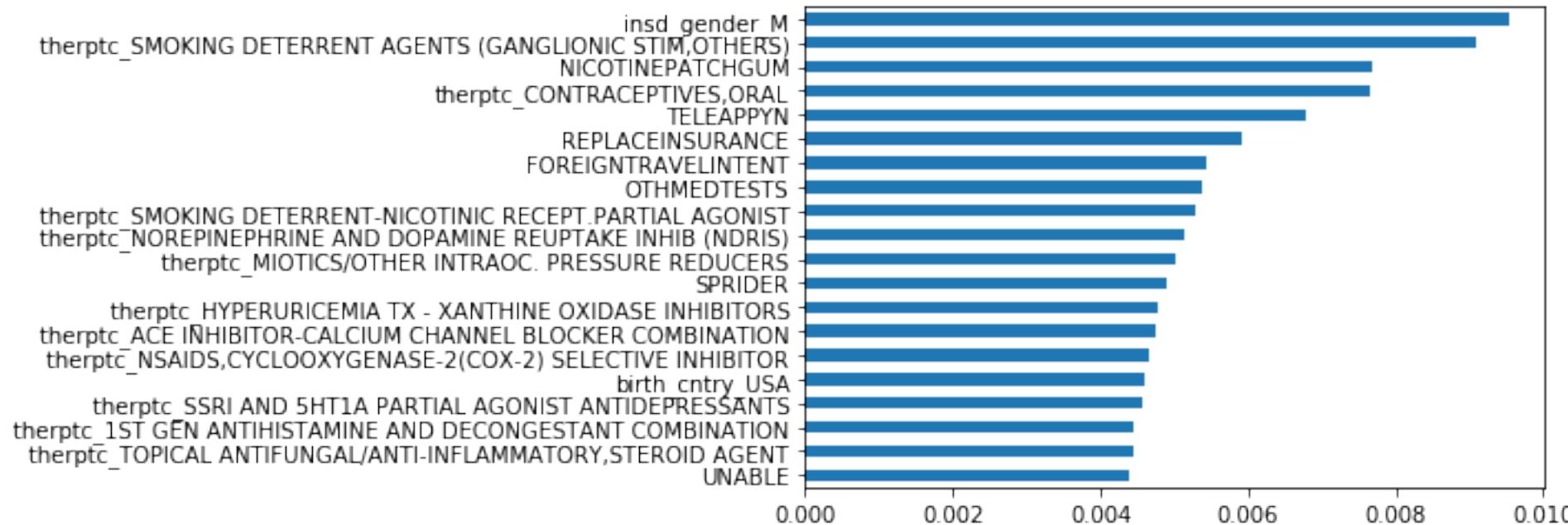
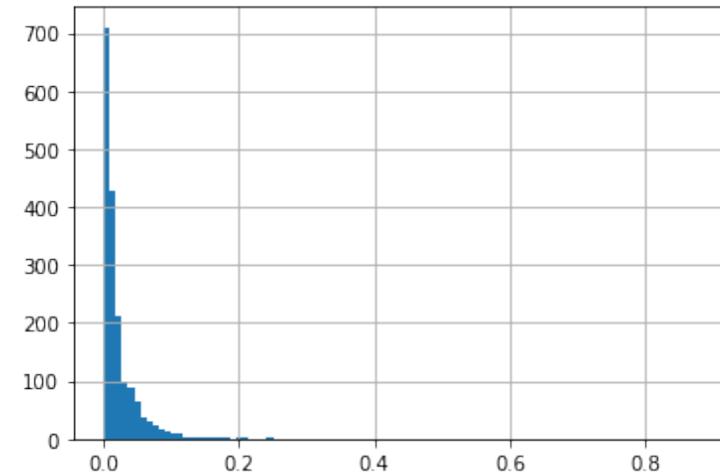
SMOKER PROPENSITY MODEL

Threshold	Precision	Recall	Accuracy
0.1	0.21	0.44	0.7479
0.2	0.38	0.28	0.8857
0.3	0.48	0.17	0.9130
0.32	0.50	0.16	0.9140
0.4	0.62	0.12	0.9170
0.5	0.62	0.10	0.9185



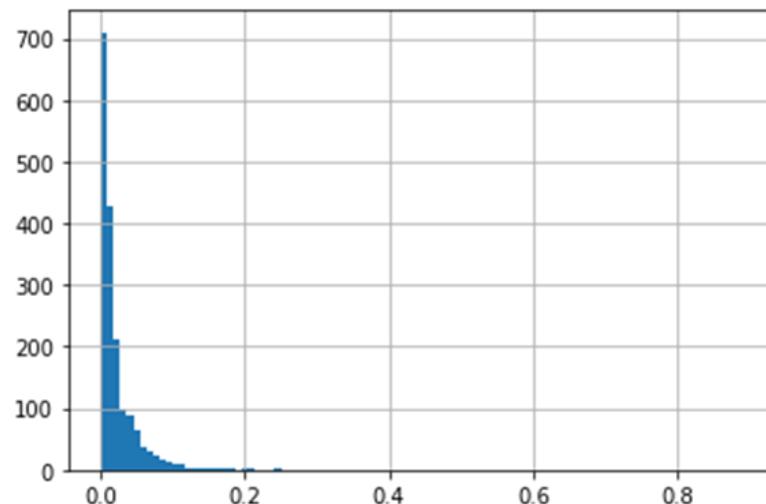
SMOKER LIAR MODEL

AUC score: 0.73



SMOKER LIAR MODEL

Threshold	Precision	Recall	Accuracy
0.1	0.21	0.16	0.9362
0.2	0.31	0.07	0.9597
0.3	0.60	0.04	0.9614
0.35	0.75	0.04	0.9614
0.38	1.00	0.04	0.9614
0.5	1.00	0.03	0.9614



COUNTRY Financial vs. Verisk vs. MassMutual

MassMutual:

"The preprocessed dataset includes ~232,000 applicants, among whom ~8% are smokers. There are 345 predictors, most of which are categorical variables."

*"Using Multi-Layer Perceptron in H2O, the 10-fold cross-validated **AUC is 0.70**.*

Verisk:

"...10,000 records..."

*"Currently for the tobacco/smoker model we have achieved an **80%+ accuracy rate** in our test results..."*

COUNTRY Financial:

- *~ 72,000 applicants, ~10% are smokers*
- *AUC is 0.74*
- *Accuracy score is 88%+*

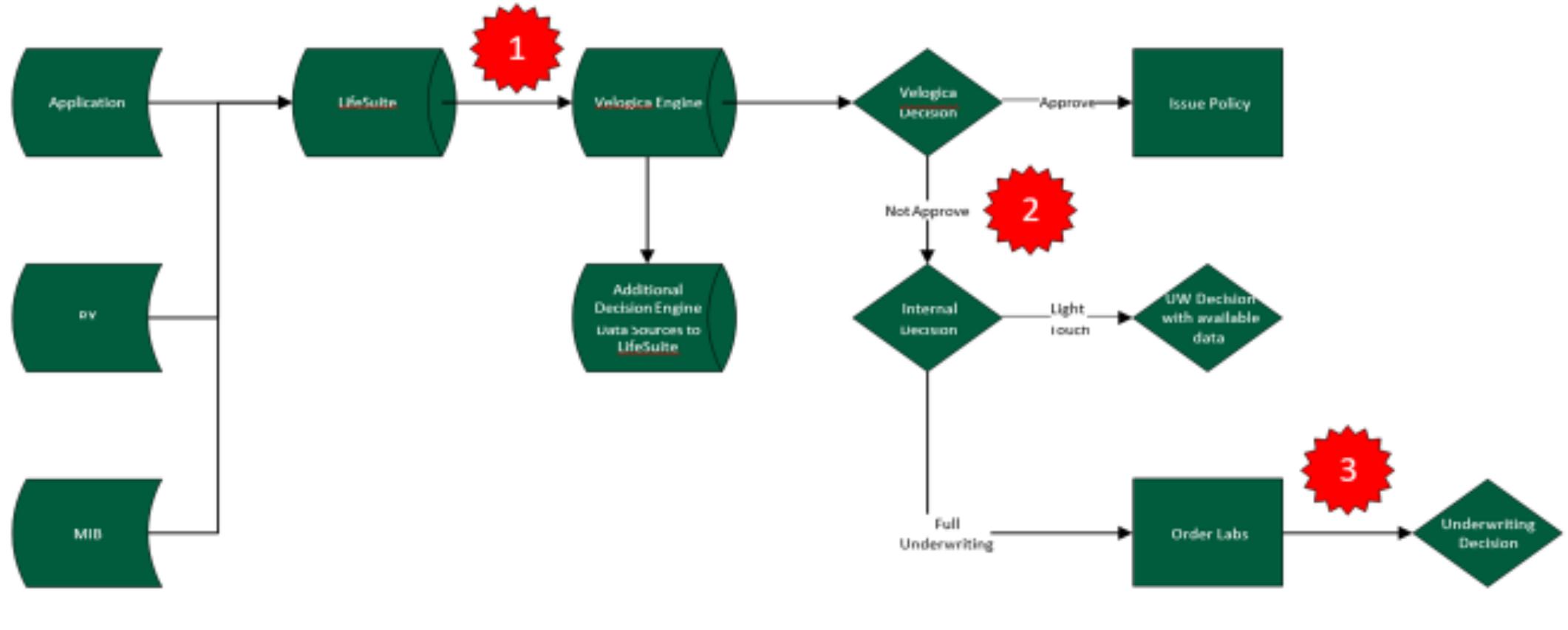
Use Cases

Smoker Propensity Model

1. Identify smokers to send into full underwriting path automatically
2. Identify applicants for which we would want to order labs after Velogica non-approval

Smoker liar model

3. Audit predictive value of model for future implementation by comparing lab results to applicants that did not admit to smoking on application. This would be in addition to the AUW random holdout set.



Next Steps

- ▶ Socialize with underwriting and actuarial team for feedback
 - ▶ Generating use cases
 - ▶ Feature engineering brainstorming
- ▶ Determine path for deployment

Thank You!