A Machine Learning Approach for Balancing Risk and Time For Evaluation of Import Entry Filers

Revati Rajane[1], Eckart Bindewald[1], Lija Fellows[2], Indu Konduri[2], Tyler Poppenwimer[2], John Jackson[2], Serajus Salaheen[2], Robert Timmons[1], Isaac Garcia[1], Brandon Tao[1], Ben Duan[1], Faiad Rahaman[2]

1: Precise Software Solutions, Rockville, MD; 2: Office of Regulatory Affairs, FDA, Silver Spring, MD

In collaboration with

Abstract

Entry filers (also called customs brokers) play an important role in the process of US food imports. They undergo evaluations by the FDA Import Filer Program. Here we present a data-driven approach that uses machine learning for risk estimation and an innovative operations research approach for balancing the risk of evaluation outcome with the time since the last evaluation. A data dashboard is augmenting this approach and provides FDA inspectors with the needed information via a user-friendly interactive display.

Filer Evaluation Program

- Entry Filer submits required product information
- Determines the quality & accuracy of the import's entry data
- Classification of the filer evaluation outcome is based on the significance of the discrepancies between submitted data and entry documents
- Routine filer evaluations are scheduled based on risk and resources
- 4-year rotational schedule in the past
- Outcomes are available to the public

Machine Learning Methodology

- Evaluation prediction: 2-class classification of acceptable versus unacceptable evaluation outcome
  "Acceptable": NAI, VAI, VWR, P1, P2, RP
  "Unacceptable": OAI, CA, SC, OC1, RU, UN
- Risk scores, past evaluations, importer history & updates and corrections to product information are considered
- Non-numeric categorical data has been encoded using mean encoding
- Balancing predicted risk with time: Combined predicted evaluation risk with time span between current date and last evaluation
- Training on data from 2,268 different import filers
- Binary classification of predicted evaluation outcomes
- Choice of lightGBM as classification algorithm - algorithm is optimizing large number of decision trees
- Classification based on multiple weighted decision trees
- Performance: Area Under the Curve (AUC) of 0.81

Eval Result	Category	Outcome
NAI	1	1
VAI	2	0
OAI	3	1

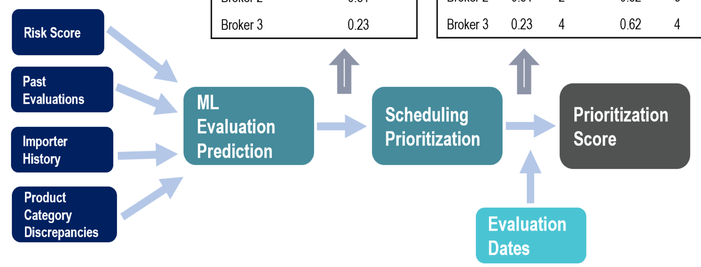
One-hot Encoding			
NAI	VAI	OAI	Outcome
1	0	0	1
0	1	0	0
0	0	1	1

Eval Result Mean	Outcome
0.037	1
0.163	0
0.379	1

Data Pipeline

Firm	Risk
Broker 5	0.44
Broker 1	0.69
Broker 2	0.81
Broker 3	0.23

Firm	Risk	Time [years]	Score	Priority
Broker 1	0.69	5	0.98	1
Broker 6	0.71	3	0.93	2
Broker 2	0.81	2	0.82	3
Broker 3	0.23	4	0.62	4



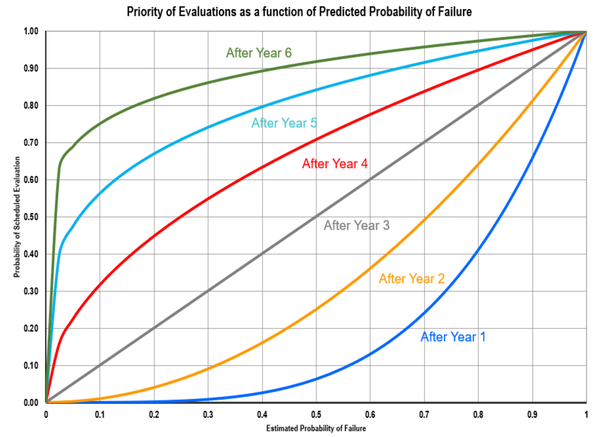
Sample ML Data

Description	Example	
Evaluation Result Code	0	History
Result of most recent evaluation- Classification*	P2*	
Evaluation type - routine, follow-up, for-cause*	RT*	
Time between 2 most recent evaluations	215	
Time between 2 evaluations prior to most recent	226	
Evaluation Result, Classification prior to most recent*	P2*	
Min percentile rank of rule-based score	66.67	Risk Percentile & Import Rule Stats
Max percentile rank of rule-based score	82.67	
75th percentile rank of rule-based score	78	
Lines that passed automated queries	17	
Lines that failed automated queries	4.33	
Total disclaimed lines transmitted by the filer	11	
Total Entries transmitted by the filer	8.33	
Country of shipment corrected by FDA staff	0	No. of Corrections by FDA
Product code corrected by FDA staff	3	
Correction on production/source	0	
Corrections on Shipper data	4	
Manufacturer data element corrected by FDA staff	0	
Quantity corrected by FDA staff	7	
Intended Use Code (IUC) corrected by FDA staff	0	
Corrected Description corrected by FDA staff	2	
Filer Fault	1	Corrections
Importer Fault	0	
No ones fault	0	
Inconclusive - Fault Non-assigned	0	
Value data element corrected by FDA staff	0	

* converted using mean encoding

Balancing Risk vs. Time

- Prioritization score is a function of time past the previous evaluation
- High-risk entities have a non-negligible probability of being evaluated if short time has passed
- Prioritization score is high for both high and low-risk entities for larger time spans



Data Visualization

- Developed an interactive data-dashboard
- Displays map view, data and accomplished evaluations in percentage
- Graphic display of information about evaluations

Accomplishments

- Developed a novel machine learning approach for estimating risk of unacceptable evaluation outcomes
- Developed a system for balancing risk vs time since last evaluation
- Provided a balanced multi-factor approach for scheduling evaluations & evaluations
- Developed an intuitive user-friendly data dashboard
- Potential applications for other similar programs

Next Steps

- Incorporation of extended filer histories
- Use of ensemble of additional classifier algorithms like neural networks, support vector machines, etc.