NRC·CNRC Data Analytics Centre

Temporary Resident applications Volume Management:

NRC Evaluation

Stéphane Tremblay

July 12th, 2018

National Research Council Canada | Conseil national de recherches Canada

Canada

Contents

Executive Summary

La méthodologie de l'initiative est excellente et suit les étapes nécessaires au succès d'un projet d'apprentissage machine. L'initiative est très bien adaptée aux différents risques organisationnels (légales, perception du public, sécurité) tout en maximisant les mesures de performances, c'est-à-dire la transparence de la méthodologie, l'efficacité des opérations et la qualité des données.

L'analyse fourni 9 recommandations et 3 projets de collaborations avec le CNRC. Les sujets abordés vont des modèles d'apprentissage automatique, de l'environnement de travail, des algorithmes ainsi que de reproductibilité des résultats.

1. Objectives

Immigration, Refugees and Citizenship Canada (IRCC) requires the services of NRC to conduct a scientific/methodological review of the tools and approaches that are used for the deployment of predictive models that will be use as a solution to support TRVM initiative.

The rational for conducting for the undertaking of a scientific/methodological review by a third independent party is to ensure both the integrity, robustness and quality of the work that had been accomplished so far.

In principle, a cursory evaluation of key phases such as Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment. Since the project is relatively nascent the focus will be restricted as described in the "Scope of Work" below.

Once the cursory review is completed, both participants will determine whether a more detailed evaluation would be required.

2. Background

The Advanced Analytics Lab (AAL) is responsible to undertake the TRVM-AA project to support the Operations Sector in achieving client service and integrity risk management outcomes in several lines of business described below.

At a high level, the approved scope of the project includes:

1. Deployment of predictive models for:
   - India and China Temporary Resident Visa applicants (eApp only)
   - India and China Study Permits (eApp only)
   - Citizenship Risk Triage
   - Passport Risk Triage
2. Development of forecasting models
3. Development of a Business Intelligence and Exploratory Environment

3. Requirement

Scientific/methodological review is essentially a "peer" review process of applying expert knowledge of acceptable criteria to determine whether the tools and protocol are adequate.

A high level report will describes what was accomplished so far including an assessment of the overall merits and identification of the tools and protocols that have been put in place.

Scientific/methodological review is a constructive process. The intent of the review is to assist IRCC to meet a minimally adequate set of criteria for the work that have been accomplished.

4. Scope of Work

The requirement is to engage the services of NRC to supply professional services expertise on the following objectives:

1. Review and assess key phases of the work undertaken so far in the following areas :
    a. Data understanding in particular on data quality;
    b. Data preparation: data cleaning technics and results, derived attributes and generated records, integration data approach and aggregations;
    c. Modelling: modelling technique and assumptions, test design, build model (settings, models, descriptions), assessment and revise parameter settings;
    d. Evaluation (efficiency of the model, deficiencies).

5. Tasks

The tasks required of NRC will include the following:

1. Develop action plan outline to conduct the review
    a. Review documentation
    b. Meet key stakeholders
2. Deliverables and Acceptance Criteria:
    a. The Plan
    b. An interim Report
    c. Final Report on the detailed and comprehensive review of the method.

6. Activities

1. Two one-day meetings took place at Immigration, Refugees and Citizenship Canada on March 20th and 27th respectively.
2. A plan and an interim reports were sent following each of those meetings
3. This is the final report

7. Analysis

La méthodologie de l'initiative est excellente et suit les étapes nécessaires au succès d'un projet d'apprentissage machine. L'initiative est très bien adaptée aux différents risques organisationnels (légales, perception du public, sécurité) tout en maximisant les mesures de performances, c'est-à-dire la transparence de la méthodologie, l'efficacité des opérations et la qualité des données.

L'analyse a également vérifiée les modèles d'apprentissage automatique, l'environnement de travail, les algorithmes utilisés afin de fournir des recommandations. Une attention particulière a été mise sur la reproductibilité des résultats.

Avec l'exception d'une activité redondante de surajustement (over-fitting), l'environnement, l'approche et les algorithmes utilisés semblent être adéquats pour répondre au besoin du projet et sa mise en production. Dans les itérations futures, une attention plus équilibrée de la variabilité du modèle et le biais serait bénéfique. L'approche utilisée dans son ensemble pour garantir la reproductibilité des résultats est excellente, simple et claire.

7.1. Environnement

L'environnement SPSS est utilisé pour tout le traitement, la modélisation et le déploiement. Cet environnement est stable et bien supporté. Son désavantage est son manque de flexibilité; les options sont limitées à ce que SPSS offre.

7.2. Cadre

L'enjeu est de classifier des demandes de visas pour la Chine. Il y a 3 ans de données étiquetées (approuvée ou non). L'entraînement des données se fait à l'écart (batch) et la performance est mesurée à l'aide de la précision de la prédiction des visas approuvés. L'approche utilisée suit les normes utilisées dans le domaine de l'apprentissage automatique.

7.3. Petite population (débalancement)

Le nombre de Visa refusé est relativement petit par rapport à celui des Visas approuvés. Les techniques de sous-échantillonnage utilisées pour réduire cet enjeu sont excellentes. Étant donné son impact réducteur sur la taille de l'échantillon, il serait utile de mieux comprendre l'effet de ce débalancement. Seulement une approche a été testée.

7.4. Modélisation

La modélisation utilise 3 années de données. La partie entraînement est composée d'applications des 3 années. La partie cross-validation (appelé testage) est composée d'application de la dernière année seulement. Et la partie testage (appelé validation) est composée de mois ne faisant ni parti entraînement ou cross-validation.

La technique de modélisation utilisée pour ce projet est un arbre décisionnel (decision tree) qui est un excellent outil reconnu pour sa transparence mais relativement moins précis que d'autres. En revanche quelques techniques de modélisation pourraient être améliorées et d'autres annulées. Une meilleure représentativité de la partie cross-validation à celle d'entraînement pourrait donner de meilleurs résultats. L'utilisation de chiffres aléatoires (seed) pour déterminer le meilleur modèle est inutile et met une pression inutile sur le over-fiting. Il semble que les données plus vieilles ont moins de valeurs que les plus récentes. Ces deux enjeux montrent que les actions liées à la modélisation sont de nature à réduire le biais au profit de la variance du modèle.

7.5. Réentraînement

La décision de réentraîner les modèles tous les trois mois est excellente.

7.6. Reproductibilité

L'utilisation du logiciel SPSS, la documentation et la justification de chaque variable, l'utilisation de « seed », la transparence du modèle choisi sont excellents pour une reproductibilité des résultats et un transfert des connaissances. Le risque d'enjeux liés à la reproductibilité est réduit à son minimum.

8. Recommendations

Suite aux deux jours de consultations, neuf recommandations ont été identifiées.

1. Avoir une stratégie de surveillance sur le réapprentissage et mise à jour des paramètres au besoin
2. Modifier l'unité d'analyse à « famille » au lieu de « personne »
3. Uniformiser l'apprentissage au groupe « Validation » (aussi appelé Testing dans la littérature) pour que les données entraîner soit semblables aux données tests.
4. Dans un environnement parallèle, développer des modèles plus sophistiqué en commençant par : Random Forest, SVM, Neural network, Deep learning, etc.
5. Augmenter, voir doubler ou tripler, la taille du groupe training (approuvé) même si le nombre de refus est relativement petit.
6. Augmenter l'échantillon en combinant les données de d'autres pays d'origine.

7. Utilisation du « clustering » pour mieux comprendre les différences entre les saisons, les pays, etc. et l'analyse de la fraude.
8. Considérer l'utilisation de logiciels libres tels R ou Python pour expérimenter de nouvelles approches et avoir plus de flexibilité sur la modélisation
9. Expérimenter les différentes formes de débalancement. Identifier l'approche qui optimise la précision des résultats.

9. Future Collaborations

Différentes opportunités de collaboration entre le IRCC et le CNRC ont également été abordées:

1. Saisir l'information capturée sur des images et l'apparier à une base de données (passeport et visa)
2. Utiliser les plateformes de calculs du CNRC pour tester des modèles plus complexes.
3. Développer du codes pour des modèles plus sophistiqués (en R ou Python par exemple)