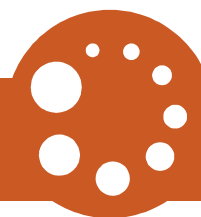




Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten

Zusammenfassung

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Herausgeber: Bundesamt für Statistik (BFS)
Redaktion: Dr. Christian Ruiz
Inhalt: Dr. Christine Ammann Tschopp, Elisabeth Kuhn, Laurent Inversin, Dr. Mehmet Aksözen, Stefan Rüber
Themenbereich: 00 Statistische Grundlagen
Originaltext: Deutsch
Übersetzung: Sprachdienste BFS

Layoutkonzept: Sektion DIAM
Download: www.statistik.ch
Copyright: BFS, Neuchâtel 2019
Wiedergabe unter Angabe der Quelle
für nichtkommerzielle Nutzung gestattet



Pilotprojekt Plausi++: Zusammenfassung

30. Juni 2019

Plausibilitätsprüfungen («Plausis») werden durchgeführt, um die Qualität und Zuverlässigkeit von Administrativdaten und Umfragedaten zu überprüfen. Daten, die dabei entweder eindeutig falsch oder zumindest fraglich erscheinen, werden an die Datenlieferanten mit Bitte um Korrektur oder Kommentar zurückgeschickt. Bisher wurden solche Plausibilitätsprüfungen entweder durch automatische Regeln und Grenzwerte oder manuell durchgeführt. In diesem Pilotprojekt («Plausi++») haben wir eine neue Art von Plausibilitätsprüfungen basierend auf «Machine Learning» (ML) Algorithmen erstellt, welche automatisch mögliche Fehlermuster entdecken kann, und sie mit einem automatisierten Feedback-Mechanismus kombiniert, mit dem eine menschlich verständliche Erklärung an den Datenlieferanten zurückgesendet werden kann. Somit konnten wir die bestehenden Plausibilitätsprüfungen ergänzen, die benötigte Zeitdauer reduzieren und die Datenqualität erhöhen.

Resultate

Ziel 1 ML-Algorithmus («prädictiver Teil»): Die Benutzung des ML-Algorithmus konnte die Datenqualität deutlich verbessern: Im Vergleich zur aktuellen Plausi (Baseline) konnte die Plausi++ **29% mehr Fehler entdecken**. Der ML-Algorithmus funktioniert mit einer **Vorhersagegenauigkeit von über 94%**.

Es handelt sich bei der Plausi++ **ausschliesslich um inhaltliche Fehler**. Die Plausi++ **ergänzt somit** die aktuelle Plausi, bei der nur 21% der Regeln inhaltlicher Natur sind und 79% **formeller Art** (z. B. Einhaltung von Codes, keine leeren Datenfelder). **Die Plausi++ ersetzt somit nicht die aktuelle Plausi, da eine Symbiose beider Plausis sowohl inhaltliche wie auch formelle Fehler abdeckt.**

Ein weiteres Indiz für die erhöhte Datenqualität ist **die Ermittlung von systematischen Fehlern auf aggregierter Ebene**. **Die Plausi++ findet insgesamt mehr systematische Fehler**. Zudem wurden beispielsweise in einer Anwendung **doppelt so viele** solcher systematischen Fehler gefunden als mit der alten Plausi.

Es gibt im «supervised ML» (überwachtes ML) zwei Phasen. Während die erste Phase viel Zeit benötigt, um den Algorithmus zu trainieren, braucht die zweite Phase von **Plausi++**, die Anwendungsphase, **nur wenige Minuten**, um ausgeführt zu werden. Zusätzlich müsste sich somit auch ein Zeitersparnis durch die Reduktion der «Hin und Her» mit den Datenlieferanten ergeben, was aber nur nach Implementation in der Produktionsumgebung getestet werden kann.

Ziel 2 Feedback-Mechanismus («explikativer Teil»): Ein automatisierter Feedback-Mechanismus erklärt das Resultat des ML-Algorithmus einem Datenlieferanten auf eine menschlich nachvollziehbare Weise. Eine solche Antwort, welche auf die mögliche Problemvariable hinweist (z. B. Alter oder Personalkategorie), konnte mit einem Verfahren vollständig implementiert werden. Weitere Ansätze könnten noch nicht getestet werden.

Ziel 3 Dokumentation: Es wurde eine detaillierte Dokumentation mit dem Ziel einer Skalierbarkeit geschrieben, damit andere Bereiche und Sektionen diese nachvollziehen und adaptieren können.

Ziel 4 Produktionsintegration: Das Pilotprojekt hat die Machbarkeit einer Produktionsintegration in



den möglichen Grenzen ausgewertet. Es gibt natürlich noch viele Unbekannte wie die schlussendliche IT-Infrastruktur-Lösung und andere Elemente, die nicht Bestandteil des Pilotprojekts waren. Wir haben dennoch Überlegungen aufgezeichnet und denken, dass die Plausi++ gut in den Prozess der alten Plausi integriert werden kann.

Fazit

Das Pilotprojekt hat alle vier Ziele erreicht und ist nach einigen Folgetests bereit in die Produktionsintegration integriert zu werden. Die Integration eines ML-Algorithmus *stricto sensu* war relativ einfach einzuführen aber die Feedback-Mechanismen schwieriger. Es liegt eine ausführliche Dokumentation für die Skalierung vor.

Verwendete komplementäre Analysemethoden:

- Für den prädiktiven Teil haben wir Algorithmen wie «Gradient Boosting Machines» und «Random Forests» verwendet, wobei wir mit ersterem eine höhere Vorhersagegenauigkeit erzielt haben.
- Im explikativen Teil haben wir «*Ceteris Paribus*» und «Local Interpretable Model-Agnostic Explanations» (LIME) verwendet. Dabei haben wir momentan den Feedback-Mechanismus mit *Ceteris Paribus* eingerichtet.