

ABSTRACT

This Master’s thesis addresses the challenge of runtime prediction for CAD data migration using machine learning techniques. The research pursues two primary objectives: predicting the migration runtime either in parts or as a whole, and conducting a comprehensive analysis of feature importance. The dataset was gathered during a specific training migration and relies heavily on metadata, exhibiting a severely right-skewed target variable. A central contribution of this work is mitigating the effects of this skewness through a novel two-stage cascading model stack. This model stack consists of Stage 1 for outlier classification, followed by outlier regression in Stage 2a and non-outlier multi-class classification in Stage 2b. The inherent class imbalance introduced in Stage 1 was successfully addressed using cost-sensitive learning techniques.

Across all stages, XGBoost consistently outperformed other algorithms, demonstrating its robustness and versatility for diverse predictive objectives. Tree-based importance thresholds and forward sequential feature selection effectively reduced dataset dimensionality while maintaining or improving model performance. A comprehensive evaluation reveals that high-level metadata, such as the initial file size, assembly type, and historical runtimes, play a significantly more crucial role than geometrical details, such as the number of structural dependencies. Furthermore, injecting domain knowledge via feature creation from log files and table columns substantially enhanced the predictive strength of the models.

The proposed two-stage cascading stack, combined with rigorous feature analysis, provides a robust methodology for addressing the research questions. Nevertheless, certain limitations remain. Predictions within individual stages still exhibit challenges, particularly regarding non-outliers. The inclusion or exclusion of correlated features did not yield a universally superior approach. Instead, it proved to be highly context-dependent while improving specific tasks like outlier prediction on customer data, simultaneously exposing a pronounced seed sensitivity due to the limited number of samples. Although the cascading approach significantly reduced the impact of heteroscedasticity, it could not completely eliminate its underlying effects. The study thus confirms the potential of machine learning methods for runtime prediction for CAD data migration and provides a methodological proof of concept.

ZUSAMMENFASSUNG

Diese Masterarbeit befasst sich mit der Laufzeitvorhersage für die Migration von CAD-Daten mit Hilfe von Machine-Learning-Methoden. Die Arbeit verfolgt zwei Hauptziele: Die Vorhersage der Migrationslaufzeit, sowohl in Teilschritten als auch als Ganzes und die Durchführung einer umfassenden Feature Analyse. Der Datensatz wurde während einer spezifischen Trainingsmigration erhoben und erfasst hauptsächlich Metadaten. Die resultierende Zielvariable weist eine extreme Rechtsschiefe auf.

Ein zentraler Beitrag dieser Arbeit ist die Abschwächung der Auswirkungen dieser Schiefe durch einen zweistufigen kaskadierenden Model-Stack. Dieser Model-Stack besteht aus Stage 1 zur Ausreißer-Klassifikation, gefolgt von einer Ausreißer-Regression in Stage 2a und einer Multi-Class-Klassifikation der Nicht-Ausreißer in Stage 2b. In Stage 1 entsteht durch die Klassifikation eine class-imbalance, die mit Hilfe kostensensitiver Lernmethoden (Cost-Sensitive Learning) erfolgreich adressiert wurde.

In allen Stages hat XGBoost andere Algorithmen übertroffen und stellt so die Vorteile von Robustheit sowie Vielseitigkeit für unterschiedliche Vorhersageziele heraus. Baumbasierte Schwellenwerte (Importance thresholds) für die Feature-Importance und Forward Sequential Feature Selection reduzierten effektiv die Dimensionalität des Datensatzes, während die Modellleistung beibehalten oder sogar verbessert wurde. Die umfassende Analyse zeigt, dass High-Level-Metadaten wie zum Beispiel die ursprüngliche Dateigröße, der Baugruppentyp und vorherige Laufzeiten eine deutlich entscheidendere Rolle spielen als geometrische Details, wie etwa die Anzahl von assembly dependencies. Darüber hinaus verbesserte das Erfassen von Domänenwissen durch Feature Creation aus Log-Dateien und Spalten der Tabelle, die Performance der Modelle erheblich.

Der vorgeschlagene zweistufige kaskadierende Stack bietet in Kombination mit der umfassenden Feature-Analyse eine robuste Methodik zur Beantwortung der Forschungsfragen. Dennoch bleiben Herausforderungen bestehen. Die Vorhersagen innerhalb der einzelnen Stufen weisen weiterhin Verbesserungspotenzial auf, insbesondere im Hinblick auf die Nicht-Ausreißer. Das Einbeziehen bzw. nicht-Einbeziehen korrelierter Features führte nicht zu deutlichen Unterschieden im Ergebnis. Vielmehr erwies sich die Resultate als stark kontext- und seed-abhängig, beispielsweise bei der Ausreißervorhersage von Kundendaten. Durch die begrenzte Stichprobengröße konnte die seed-sensitivity nicht weiter eingegrenzt werden. Zwar konnte der kaskadierende Ansatz die Auswirkungen der Heteroskedastizität signifikant reduzieren, doch deren zugrunde liegende Effekte nicht vollständig eliminieren. Die Arbeit zeigt damit das Potenzial von Machine-Learning-Methoden zur Laufzeitvorhersage von CAD-Daten Migrationen und liefert einen methodischen proof-of-concept.