Reducing Receipt Mismatch Costs Using Categorical Variable Imputation

Aniket Banerjee[1], Meera Govindan[2], Shubhansh Jain[3], Sushil Achamwad[4], Daniel Lee Whitenack[5]

Purdue University Krannert School of Management

banerj33@purdue.edu[1], govindam@purdue.edu[2], jain297@purdue.edu[3] sachamwa@purdue.edu[4], dwhitena@purdue.edu[5]

Abstract

This study investigates the likelihood of mismatch in receipt and invoice matching processes for account payables, called 'receipt grief' for an American Fortune 100 company, which is also the world's largest construction equipment manufacturer. The business unit incurs an annual economic loss of USD 6 million for the rectification of receipt grief. Our study demonstrates how high receipt grief costs can be reduced without manual intervention via a model that predicts the likelihood of receipt grief. We have used R & Python (PyTorch, Scikit-learn) for developing a predictive model to identify the possibility of receipt grief & visualization tools (Tableau) to generate insights.

Introduction

- The company's Payables department is responsible for receiving, processing, and approving more than 100K carrier invoices per month containing around 320K invoice lines
- The otherwise automated process requires human intervention when the payables application is unable to match invoices to the corresponding receipt record, halting the process
- The break down in the process is known as grief. When grief occurs, the invoice is assigned to an analyst for resolving the grief by manual verification
- Grief is time consuming and costly. The likelihood prediction of resulting grief is based on invoice attributes including supplier codes, invoice number, locations, charges, and receiving facilities
- If the likelihood of grief is sufficiently large, an invoice can be sent back to the receiving facility for review, thus saving manual intervention costs for the Payables department

Literature Review

Study	Motivation/Methods Described
Multiple Imputation by Chained Equations: What is it and how does it work?	Many researchers have not been trained in the MICE method and few practical resources exist to guide researchers on the implementation of this technique. This paper introduces the MICE method with a focus on practical aspects and challenges in using this method.
K- Nearest Neighbor in Missing Data Imputation	The paper proposes a comparative study on single imputation techniques such as Mean, Median, and Standard Deviation combined with K-NN algorithm.
Reengineer the payables process	In the current competitive business environment, companies are aggressively cutting costs by reengineering processes, and the idea of this paper is to focus the attention on increasing the efficiency of the Accounts Payable process.

Methodology

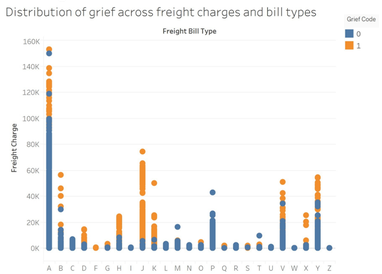
Study Design

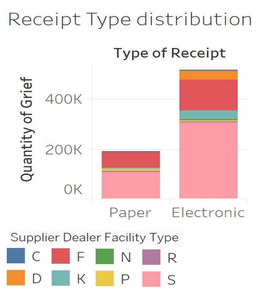
Data

- American Fortune 100 industrial machinery company's invoice-receipt grief data for 2011 (Q1-Q4) - 2012 (Q1-Q4)
- Around 4 million rows of categorical data with more than 700K records that have receipt grief

Data Cleaning & Pre-Processing

- Exploratory Data Analysis used to remove anomalies and outliers
- Caret package was used extensively
- Categorical variables were converted to dummies





Feature Engineering

- **Normalization:** Numerical predictors – Freight Weight and Freight Charge using Min-Max normalization
- **Category Binning:** Categorical predictors binned using frequency distribution due to abstract domain understanding and lost meaning

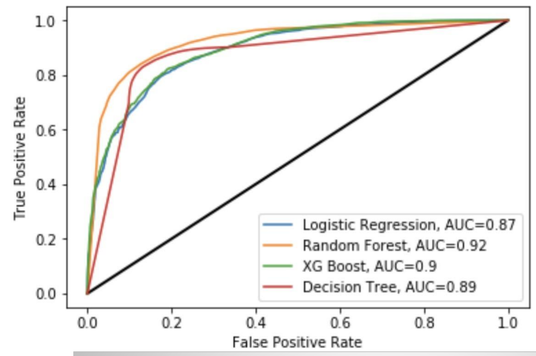
Model Design and Methodology Selection

- Data split into 80:20 partitions for training and testing data
- Attempted Modelling Methods:
  Logistic Regression, Decision Trees, Random Forest and XGBoost
- Above models were chosen depending on various factors such as variance, bias, input features independence

Model Evaluation / Statistical & Business Performance Measures

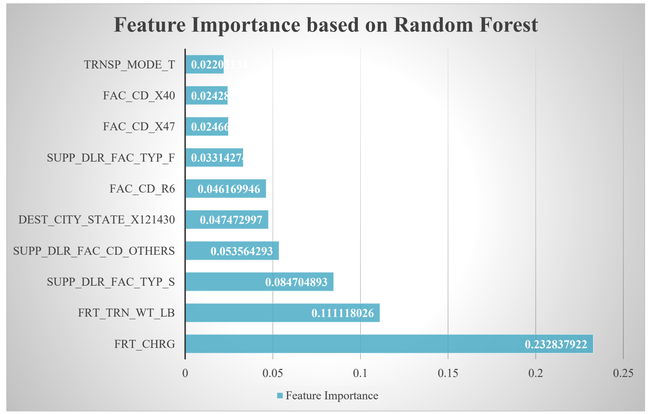
Models were evaluated on Accuracy (ACC), Sensitivity, Specificity and Receiver Operating Characteristic (ROC) Curve

Results



The ROC curve suggest the best likelihood prediction using the Random Forest model on the test data set.

Metric	Performance
Accuracy	0.87
Specificity	0.61
Sensitivity	0.96
AUC score	0.92



Using our current best model, we can reduce the losses in the form of receipt grief by more than 80%. This can potentially result in an annual savings of almost USD 5 millions to the company.

Conclusions

Our business problem revolves around the economic losses of receipt grief borne by the client. With prediction accuracy of 87%, the random forest model is an ideal candidate for understanding the critical factors affecting receipt grief and hence can be used in invoicing systems to predict the discrepancies in the invoicing processes. The predictions will also empower stakeholders to make informed decisions pertaining to future economic gains and losses for the company.

Future Scope

- With knowledge of the business domain, data can be interpreted more meaningfully for category binning and model creation
- The on-going work has a promise of increasing prediction efficiency to more than 90%, resulting in further cost savings