Predicting Future Warranty Costs

INTRODUCTION
About Wabash:
Wabash is an American industrial manufacturing company of engineered solutions and services for transportation, logistics and distribution industries, as well as North America's largest producer of semi-trailers and liquid transportation systems.

Our goal:
• Build a model to predict future warranty costs for Wabash
• Utilize historical warranty, configuration data, and product specifications to gain meaningful insights to forecast warranty costs
• Investigate the fail codes and the factors that attribute to specific failures to understand warranty trends and root causes
• Analyze the different predictors that will influence warranty costs

Why:
• More efficient planning for future warranty costs
• Improve product quality
• Strengthen value proposition for new products

EXPLORATORY DATA ANALYSIS (EDA)

MODEL BUILDING
Our models:
• Split the data into training and testing sets and sees how well the test set performs
• Provides an accuracy value based on how well it fits the test set
K-Nearest Neighbors
• Used different values of k in order to optimize accuracy
• Highest accuracy obtained was 68% using k=20
Logistical Regression
• Used configurations that were correlated with fail codes to predict likelihood of a specific fail code occurring
• Coefficients were determined by strength of correlation; likelihood calculated using binary configuration data
Decision Tree
• Test that closely mirrors human decision making and can handle the binary configuration data easily without creating dummy variables
Random Forest
• Same as a decision tree except for a small tweak that decorrelates the trees
• Results did not significantly increase accuracy compared to the decision tree

REDUCING FEATURES
• We found that numerous features had an insignificant number of appearances in claims in the data
• We also found that some features were always present alongside one another so we merged these features to avoid collinearity in the predictive models.

CHALLENGES/ROADBLOCKS
Model Confusion
• We initially tested a variety of regression models using qualitative predictors (fail code, trailer count, etc.)
• Very low R^2 values (see figure below) indicated poor linear fit, leading to confusion on which models to implement
• Eventually obtained configuration data as new predictors

Data Flaws
• Inconsistent attribute data led to issues with the training model and caused potentially key information to be left out of analysis
• Our dataset contained columns of free text meaning to be identical, which led to misspellings/abbreviations that made the data inaccurate

CONCLUSION/FUTURE PLANS
What’s next:
• The models we built can be continuously updated by Wabash to predict their future failures and costs associated with them
• As we approach the end of the year, we plan to improve accuracy of our models by updating code, reducing more configurations, and implementing new tests such as chi-square tests and neural networks
• We also would like to summarize our findings in a Power BI dashboard

ACKNOWLEDGEMENTS
Aaron – IT Director
John – Data Analytics Manager
Evan – Data Analyst
Miku – Senior Manager of Warranty
Mark – Warranty Supervisor
Lauren – Data Mine Corporate Partners Advisor
Kali – Data Mine Associate Research Engineer
Poomin & Darsh – Wabash Project TA

DISCLAIMER: ALL NUMBERS ARE THEORETICAL