EXECUTIVE SUMMARY

The GE Transportation Practicum Team conducted research on the firm’s past history, goals as an organization, and technical methods needed for the deliverables in preparation of this practicum project. Our project goal was to use anomaly detection methods and data visualization to easily identify locomotives with non-normal maintenance behavior. Key pieces of information and key performance indicators are communicated to the user through a dashboard, containing various amounts of locomotive information and interactive visualizations. We developed a visualization dashboard for both GE Transport Analyst and Management groups, allowing GE Transportation to improve the internal efficiency of locomotive repair schedules and also enhance decision-making capability. We also performed an in-depth predictive modeling analysis that examined key variables of interest in predicting the effective life\(^1\) of a locomotive combo unit.

As part of the practicum project, the GE Transportation practicum team also has some other notable findings to report, which we outline later in this document. A key point is that our final analytics dashboard should be compared to a regression analysis. In order for an analytic measure to be included on the dashboard in must generalize well to GE transportation’s customer base. We conducted many ad-hoc analyses, which frequently produced interesting findings for small numbers of locomotives, but we remained focused on our task of producing analytics and visualizations that could be used consistently and meaningfully for GE’s customers at large.

STATEMENT OF OPPORTUNITY

GE Transportation is a market leader in the locomotive industry. GE Transportation collects data from locomotives on a regular basis that includes fault codes and other diagnostic information. In addition, GE Transportation collects repair data on ‘combo’ units from two repair shops, in Las Vegas and Kansas City.

GE Transportation is seeking to improve their internal efficiency and increase margins by using analytics to drive down cost associated with locomotive maintenance. By analyzing the data contained in from these and associated datasets, GE Transportation could increase efficiency by conducting preventative maintenance on locomotives, or by re-engineering internal processes of locomotive repair within the shop. Over the past nine months, our team has identified some key pieces of information which could improve the internal efficiency of the locomotive preventative maintenance.

\(^1\) We define the “effective life” of a combo unit as the time between when a combo unit leaves the shop and when it returns.
BENEFITS

Key performance indicators (KPIs) are communicated through our visualization dashboard, which could ease data interpretability and also broaden the audience to which data-driven insights are communicated. These KPIs are count of infancy failures, frequencies of combo shipments, PE and CE bearing temperatures, and frequencies of reason changed. In addition, our predictive modeling analysis provided some key information as to what are the most significant drivers of a combo unit’s effective life.

PRACTICUM DELIVERABLES

DATA CLEANSING & EXPLORATORY DATA ANALYSIS

We narrowed our data to a few variables from some very dirty and vast data (over 700 variables) according to GE’s recommendations, our own judgement based on context, number of non-NA observations available in each variable, and the ranking of variable importance using the initial classification and regression tree model. There were 5786 observations in total after cleaning, and 1364 complete cases; depending on the nature of various models, we used either only the complete cases or the whole dataset including NAs.

Some variables of interest
For all three variables, extensive data cleansing was needed. The general approach was to strip any and all text characters from the data, and convert the remaining values to a common numeric format.

The CE Temp and PE Temp variables required some additional interpretation and cleansing work, beyond the aforementioned cleansing process. Since both Celsius and Fahrenheit temperatures were recorded in both variables, the distributions of CE and PE Temp were plotted in order to isolate which temperatures were Celsius and which were Fahrenheit.
The two distributions of CE and PE Temp (showed above) pointed to a clear distinction between temperatures recorded in Celsius and temperatures recorded in Fahrenheit. Temperatures below 60 on this scale were taken as temperatures on the Celsius scale, which were then converted to temperatures on the Fahrenheit scale. This resulted in a common scale for all temperatures recorded in the CE and PE Temp variables. The final distributions of variables are
shown jointly below, after both variables were converted to Fahrenheit-only values.

In addition to the three variables, the PE 12 IRD variable was also cleansed and modified to a numerical variable using a highly similar approach to that used for the aforementioned variables. The distribution of this variable is shown below. Note that the distribution is left-skewed and contains values on the upper end of the distribution that are roughly equidistant from one another.

The cleansing of the allowed us to improve
the accuracy and classification rates of the models fitted.

PREDICTIVE MODELING

We used supervised machine learning models, including, but not limited to, random forests and boosted trees in our modeling which predicted the effective life of a combo unit. The importance of predicting whether an combo unit will have a short effective life is two-fold. First, GE forecasts the number of combo units needed to be replaced. By having a better estimation of when a combo on a specific locomotive is likely to fail, GE can aggregate their forecasts and perhaps keep idle inventory at a minimum. Second, by knowing what processes or measurements taken in the shop indicate a longer running combo, GE can track what contributes statistically to shorter combo timespans and avoid such efforts (and vice-versa). This saves everyone time and money: customers have less down-time on their locomotives and can conduct more commerce, and GE can save money on locomotive service contracts with less frequent or less expensive servicing.

We ran several types of models; classification and regression tree (CART), projection pursuit regression (PPR), neural networks, gradient boosted tree, random forest, logistic regression, and support vector machines. We ended up covering an entire host of model types far more extensive than what was outlined in our original proposal, in which we stated we would cover classification and regression trees (CART) and logistic regression. Overall, the model that showed the best overall performance was the gradient boosted tree. Given the messiness of our data and the low amount of complete cases, this result was not unexpected. Tree-based models handle missing data far better than the non tree-based models we evaluated here; which were logistic regression, support vector machines, neural networks, and projection pursuit regression. However, to our benefit of understanding, logistic regression includes coefficients that allow us to examine the direction of the effect and compare the magnitude of the effect of the predictors.

CLUSTERING ANALYSIS

We used k-means clustering on several occasions to isolate clusters of locomotive combos that performed well and poorly. Getting to the clustering stage often meant substantial data cleansing. Though clustering often proved to provide ad hoc insights, the data sets that we were ultimately able to condense for use in clusters were not general enough for GE’s use. For example, we ran the k-means analysis below.
The analysis was useful to our group since it forced us to think about what constitutes a poorly performing locomotive/component. Parts and locomotives wear-out. What we’re trying to isolate are repair schedules that are suboptimal as well as components that are wearing-out too quickly. In order to run an actionable clustering analysis we had to join several tables at a low merge rate, and thus we can’t claim that our results are representative of the customer population for GE locomotive combos. However, the above analysis shows tend to return to the shop with above average measurements for components parts that wear-down, and a below average period of time from when the component was in the shop. The locomotive combos are returning to the shop before they’re “worn-out.” An important caveat to mention is that the data we ran the analysis was filtered to remove maintenance. It only included component failure data.
Another good finding included that when comparing combo-lifetime or span to rim thickness, that six distinct k-means clusters optimally segment (as calculated by the pseudo-F value) a specific group of BNSF locomotives.
We liked the visualization because it gave ample opportunity to tell a story for why a specific group of locomotives was performing above or below the average and inner quartiles. The example above shows that two clusters that consistently were worked hard had problems with gear-case oil levels when they came into the shop. We ultimately decided to use another version of the above analysis that did not use k-means as a way of segmenting the clusters. Since we’re passing the dashboard off, it seemed that unless one knows R, and how to incorporate it into Tableau, then the dashboard could easily break. However, as a finding we

**VISUALIZATION DASHBOARD**

As part of our practicum project, we designed and implemented a visualization dashboard that can be used by both GE Transport Analyst and Management groups, which could improve the internal efficiency of locomotive repair schedules and also enhance decision-making capability. Details about which datasets are used, key variables of interest, our design approach, and screenshots of the final outcome are all included below.
DATASETS USED

Our visualization dataset is an aggregation and result of a merger between several key datasets provided by personnel within GE Transportation. We merged the datasets by work-order number. After the merger of the four datasets, our final dataset which we used for visualization included 15,564 unique observations. It is important to note that our dataset includes only locomotives with customer BNSF, but our general approach could be extended to data incorporating other customers. A diagram of our merging strategy is included below.

KEY PERFORMANCE INDICATORS (VARIABLES OF FOCUS)

Based on feedback provided by the GE Transportation analytics team, as well as GE Transportation engineers, our visualization dashboard includes several Key Performance Indicators (KPIs) that are of use throughout the organization. These KPIs that are integrated into the dashboard not only provide GE Transportation with a broad overview of general locomotive performance, but also with additional details on demand through the use of filters that can provide additional key insights.
COUNT OF INFANCY FAILURES

Infancy failures are the least desirable and most costly event that could occur with a locomotive, resulting in a premature visit to GE Transportation repair points (KCR or LVR), with unexpected delays and cost to both GE and GE customers. In the most extreme case, infancy failures could result in a derail or other on-track stoppage, further exacerbating the issue at hand and increasing cost above costs needed to repair the locomotive prematurely.

We included infancy failures as part of our main page, as well as other associated views.

FREQUENCIES OF COMBO SHIPMENTS

Each additional combo shipment to a repair facility, such as KCR or LVR, brings with it additional cost for GE Transportation. We included views of combo shipments, both geographical and as a heat map, for GE Transportation in our final dashboard. These views provide GE with the ability to quickly see where the majority of combo shipments occur by repair point and organization code.

FREQUENCIES OF REASONS CHANGED

Last but not least, we included frequencies of reasons changed in the final dashboard. While the vast majority of combo repairs or replacements are the result of diminished rim thickness, it may be interesting to explore other reasons changed that could point to issues with a particular locomotive, locomotive model, or repair facility. We’ve included the frequencies of reasons changed as part of our second main sheet on our dashboard.
OTHER VARIABLES OF INTEREST

RUN TIME

Serial numbers are tracked when locomotive components enter or leave a repair shop - this allows GE to keep track of traction motor run times. Run time could have been a key variable to perform additional exploratory data analysis and to validate some of the hypothesis that were formed from analyzing the aforementioned variables. However, merging the given run time and proficy datasets resulted in only 1204 usable observations (composed of both DC and AC data points). Checking obvious patterns, such as derailed locomotives having a very short run time (probably due to a serious failure occurring shortly after the locomotive left the repair shop, is possible with the resulting dataset. However, more complex correlations cannot be made reliably with such few data points. Run time was therefore omitted as a standalone KPI to be represented in the dashboard.

DESIGN APPROACH

Our design approach follows the familiar and well-respected notion that visualizations should provide broad overviews of information on initial approach, with details on demand through the use of hover, click, and zoom effects. We also incorporated the ability to filter the data according to certain key attributes, such as date range, traction motor type, repair point (ocean point), or reason changed.

Our visualization dashboard includes five main sheets: the main page, volume trends, shipment flow, bearing temperatures, and rim thicknesses. Each sheet focuses on a distinct, but related, set of information that will enable the user to gain a deeper understanding of combo repairs within GE Transportation.
The main page is the first window that a user will see when using the Tableau dashboard. This page gives the user a quick look at how well the company is performing (in terms of incoming repairs) and displays important information regarding the reason changed, the count of incoming repairs and the percentage of infancy failures in the past 28 days, 6 months or 1 year. The sunburst on the upper-left corner of the page allows the user to filter out the data, using customers, type of traction motor (DC or AC) and fleet ID as the constraining variables. Upon clicking the GE sunburst, the heatmap and line charts will be automatically updated to only display information based on the data subset of interest.
The volume trends visualizations show quantities of combos received, the proportion of combos received that are infancy failures, and the associated reasons changed for a particular time interval. Filters allow the user to filter by customer, traction motor type, motor model, and the PTN date associated with a combo repair.
This page explores combo shipments across North America by pick up point (organization code) and repair point (ocean shop), with the ability to examine shipments by fleet ID as well. The purpose of this set of visualizations is to examine shipments of combos and see if there are any organizational efficiencies that could be gained by altering supply chain of combo shipments across North America.
This page explores historical averages of PE and CE bearing temperatures by locomotive ID and allows the user to explore these values by filtering the reported defect, ocean shops (repair point), customer, traction motor type, and fleet ID. By quickly identifying which locomotives are performing below average (or above average), GE Transportation can direct resources to those locomotives for further examination. Having a better understanding of why a particular locomotive performs better or worse than average could point to broader solutions that positively impact the company and operations.
This sheet allows GE Transportation a good approximation of how groups of locomotives are performing relative to one another. It’s an extension of our previous clustering analysis, except it allows the user quick approximations of performance that are not produced from high-level statistics, and therefore allows enough generalization and use robustness to be included in the dashboard.

IMPLEMENTATION PLAN

We have included an implementation plan for the presented GE Transportation Dashboard as part of our final deliverable, which is included in the appendix of this report. The implementation plan covers R code necessary for data cleansing, code necessary for data aggregation, and how to import this data into the Tableau dashboard for easy viewing.
OTHER NOTABLE FINDINGS

Early on in the practicum, we performed exploratory analysis in order to identify fleets of locomotives that showed unusually high rates of repair. We identified BNSF fleet 66 as having unusually high rates of repair relative to other fleets within BNSF. We broke down fleet 66 and examined the fleet according to both the motor model and reason changed. While rim thicknesses are normally the primary reason of repair for any given fleet, BNSF fleet 66 shows unusually high counts of traction motor failures and also loss of lube failures.
Another notable finding that should be explored further by analysts at GE Transportation is the difference between bearing temperatures between the KCR and LVR repair facilities. While this could be explained by differences in ambient temperature between the two locations, if that is not the case, then this finding could present an interesting area of improving quality at the LVR facility.

TOOLS, LANGUAGES, AND METHODS USED

- **Statistical Computing Languages:**
  - R
  - Python
- **Data Visualization Libraries & Tools:**
  - Base R Graphics
  - ggplot2
  - Tableau 9
**Methods & Approaches:**

- **Exploratory Data Analysis**
  - Examined general trends across customers, fleets, and locomotive types
  - Identified locomotives that consumed an abnormally large amount of combo units

- **Supervised Learning:**
  - Logistic Regression
  - Classification and Regression Trees
  - Boosted Trees
  - Random Forests
  - Neural Networks
  - Projection Pursuit Regression
  - Support Vector Machines

- **Unsupervised Learning**
  - Clustering
    - K-Means Clustering
    - Gaussian Mixture Models

- **Visualization Dashboard Development**
  - R
  - Tableau 9

---

**CONCLUSION & PATH TOWARDS THE FUTURE**

We performed exploratory data analysis, predictive modeling, clustering analysis, and development of the final visualization as part of our nine-month practicum project. We believe the project has been a success, a success that will allow GE to broaden the interest of analytics within the company. As part of our final deliverables, we have included an implementation plan which is included in the appendix, in addition to the slides, final presentation, and this report. We would like to thank both Katie Tong and Sheng Li for their outstanding support during the last 9 months, as this practicum project would not have been possible without their help. We’d also like to thank the MSiA faculty, staff, professors, and lecturers as well, as they provided the technical training, fuel in the form of food, and continued exposure to fresh ideas that we used for the entirety of the practicum project.
APPENDIX

DASHBOARD IMPLEMENTATION

As part of our final deliverable, we have included all necessary R scripts, Tableau workbooks, and other data as needed in order to update the dashboard with new data and implement the data within GE Transportation. This folder is accessible on the MSiA computer cluster and will also be emailed to GE Transportation. Within the main directory, there are subdirectories that contain associated data files, R scripts, and Tableau workbooks. Data preparation for the vast majority of visualizations occurs in “data_prep1” which is contained within the R scripts subdirectory. The remainder of the data preparation occurs in the remainder of the R scripts. In all R scripts, data sources and the reference to those data sources should be updated before generating new data for the visualizations. Once those changes have been made, those interested in updating the data should simply be able to run the entirety of each R script to update the datasets used in visualizations. We believe that displaying the dashboard publically within the organization, as we show below, could have a positive effect on the company.
Once the data export is complete, simply point the Tableau workbook to the updated and refresh the data source. The updated visualizations will show after allowing Tableau to update the data source.

**TASK LIST FOR GRADING PURPOSES**

Data Cleansing (joint)
- EDA
  - Proficy (mostly Valentino, but joint)
    - Date Cleansing in Proficy (Elie and Valentino)
  - Wheels2 (joint)
  - BearingLife (joint)
  - Bull Gear vs Rim Wear (joint)
- Predictive Modeling
  - Neural Network (Elie)
  - Projection Pursuit Regression (Valentino)
  - Support Vector Machine (Tina)
  - Logistic Regression (Steven)
  - Classification and Regression Tree (Steven)
- Boosted Tree (Kapil)
- Random Forest (Kapil)

- Clustering Analysis
  - Gaussian Mixtures (Tina)
  - K-Means (Steven)

- Visualization Dashboard
  - Preliminary Visualizations (joint)
  - Final Visualization Dev (mostly Tina and Kapil, but joint effort through iterative design)
  - Final Visualization Adjustments (Tina, Kapil, Elie)

- Final Report (joint)
- Final Deliverables (joint)
- Presentations
  - 2/26 (joint)
  - 4/21 (joint)
  - 6/08 (joint)