Ross Conway & Alec Schneider

Project Checkpoint #2

5/29/2021

## Specification

Our team’s goal is to accurately forecast maintenance costs for the grocery store chains using our work order management systems (WOM). Being able to accurately forecast maintenance costs for these chains will drive proper investment into physical assets, more accurate shareholder guidance, and improved accrual accounting for the grocery store chains.

If times allows, we will attempt to find relationships between different maintenance items to see if costs can be lowered through earlier maintenance investment.

## Observation

We found that there are five major grocery chains that we will focus on forecasting for the next 12months:

* Giant Food
* Giant Martins
* Stop & Shop Central
* Stop & Shop North
* Stop & Shop South

Before looking at the time series elements of the individually grocery store data we wanted to begin understanding the distributions of some of our categorical columns. This will inform us where labor and costs are dedicated.

Number of unique values for each main categorical variable. This is important for considering whether we can one hot encode a column, or if the one hot encoding transformation will create too many columns and therefore, an extremely large dataset by size. Work orders are a unique identifier, which is why our unique values of wo# are so high.

Unique values for wo#: 324434

Unique values for property\_number: 758

Unique values for specialty: 78

Unique values for portfolio: 5

Unique values for priority: 5

Unique values for district: 47

Unique values for region: 10

Unique values for status: 2

Unique values for organization: 188

Unique values for item\_asset\_model: 364

Unique values for type\_category: 3

Unique values for type: 6

Unique values for work\_order\_city: 566

Unique values for work\_order\_state\_prov: 12

Unique values for work\_order\_zip\_postal\_code: 1313

Unique values for repair\_category: 20

Unique values for repair\_code: 71

Unique values for reason: 9

Unique values for specialty\_group: 6

Let’s check the percentage of null values these categorical variables contain. We can see that repair\_category, repair\_code, and reason will need to be dropped from the dataset as the columns are too sparse to correctly impute:

wo# 0.000

property\_number 0.000

specialty 0.000

date\_time\_created 0.000

total\_charged\_amount 0.000

portfolio 0.000

priority 0.000

district 0.003

region 0.003

date\_time\_completed\_\_last 0.092

status 0.000

organization 0.099

item\_asset 0.000

item\_asset\_model 0.000

type\_category 0.000

type 0.000

total\_cost 0.000

work\_order\_city 0.000

work\_order\_state\_prov 0.000

work\_order\_zip\_postal\_code 0.000

repair\_category 0.965

repair\_code 0.965

reason 0.902

total\_labor\_hours 0.000

created\_to\_picked\_up 0.000

created\_to\_started 0.000

total\_labor\_cost 0.000

total\_services\_cost 0.000

total\_materials\_cost 0.000

total\_spot\_purchase\_cost 0.000

total\_miscellaneous\_cost 0.000

specialty\_group 0.009

### Specialty:

Percent of Work Orders Allocated to Top 5 Specialties: 51.35%

Percent of Work Orders Allocated to Top 10 Specialties: 73.90%

Percent of Work Orders Allocated to Top 20 Specialties: 88.63%

Percent of Work Orders Allocated to Top 25 Specialties: 91.86%

Chart, histogram

Description automatically generated

Percent of Total Charged Amount Allocated to Top 5 Specialties: 65.90%

Percent of Total Charged Amount Allocated to Top 10 Specialties: 77.66%

Percent of Total Charged Amount Allocated to Top 20 Specialties: 89.20%

Percent of Total Charged Amount Allocated to Top 25 Specialties: 92.49%

Shape

Description automatically generated with medium confidence

### Geographic analysis

Most of our total charged maintenance costs lie in the Northeast region of the United States of America:

Chart

Description automatically generated with low confidence

### Maintenance Organization

Top 25 organizations hired by our grocery chains by mean labor hours: 

Chart, scatter chart

Description automatically generated

## Analysis

Maintenance Cost Forecast:

* As a baseline model we will use Facebook’s Prophet time-series forecasting package to create an additive time-series model for each grocery store chain
* ARIMA will be the next technique that we will use to make time-series predictions for each grocery store chain. The ARIMA models will be compared against the baseline Prophet models
* If we are not satisfied, we will use a deep learning architecture called a Transformer, that can model time series data by encoding and decoding past data to capture state behaviors in the underlying data. Our modeling team has a GPU to decrease model training times significantly

## Recommendation

1. First, issue we will decide upon will be the final decision regarding imputation of missing values
   1. As of now, we are planning to use specialty means to impute missing numeric data, and replace missing categorical values based on the most common items for that column by grocery chain
2. Next, we will process our creation date column into hour, day, and year cosine and sine signals to enhance the deep learning model
3. Then we will one hot encode the categorical variable of our choice to add features to the dataset
4. After all preprocessing steps are finalized, we will build a datapipline to transform our data on the fly, saving storage space on our machines
5. Once the data pipeline has been defined, we will finalize our modeling designs
6. Finally, we will analyze our individual models and evaluate backtest and future performance
   1. If performance of the models remain inadequate after all iterations, we will explore leveraging a multi-model pipeline

## Data Dictionary

* WO# - Work Order Identification Number – String (Categorical)
* Property Number – Property Identification Number – String (Categorical)
* Specialty – Type of Work Needed – String (Categorical)
* Date/Time Created – Work Order Creation Date - Timestamp
* Total Charged Amount – Total $ Amount Charged for Work – Float
* Portfolio – Portfolio Location – String (Categorical)
* Priority – Scale of Importance of the Work Order (Ordinal)
* Customer GL Account Code – Customer’s General Ledger Account Code – String (Categorical) - DROP
* Work Description – Description of the work being requested on the work order – String - DROP
* Work Order Address – Address of where the work is being performed – String - DROP
* District – District where the work is being performed – String (Categorical)
* Region – Region code of the location where the work is being performed
* Date/Time Completed – Last – Timestamp of the work order completion – Timestamp
* Status – Status of the work order (Completed/Cancelled) – String (Categorical)
  + If status == Cancelled, then the Date/Time Completed is Null
* Organization – Name of the company performing the work – String (Categorical)
* Item Asset – Item being repaired or replaced on the work order – String - DROP
* Item Asset Model – Model type of the Item Asset – String
* Type Category – Type of work order being submitted – String (Categorical)
* Type – Type of work order submitted – String (Categorical)
* Total Cost – Total $ cost of labor, materials, and services – Float
* Work Order Address 2 – Street address of the property that is requesting the work order – String – Drop
* Work Order City – City of the property that is requesting the work order – String (Categorical)
* Work Order State/Prov. - State of the property that is requesting the work order – String (Categorical)
* Work Order Zip/Postal Code – Zip code of the property that is requesting the work order – String (Categorical)
* Repair Category – Type of repair being – String – DROP
* Repair Code – Detailed repair code – STRING – DROP
* Property Name – Name of property where the work is being performed – String (Categorical) - DROP
* Reason – Reason for cancellation of the work order – String (Categorical) – DROP
* Reason Note – Details on why the reason was selected for the cancelled work order – String - DROP
* Total Labor Hours – Labor hours performed for the work order - Float
* Created to Picked Up – Time since work order created till received by vendor - Float (Need to be converted from H:M:S)
* Created to Started – Time since work order creation to starting the work – Float (Need to be converted from H:M:S)
* Total Labor Cost – Total $ cost of the labor needed to perform the work – Float
* Total Services Cost – Total $ cost the services performed to complete work order - Float
* Total Materials Cost – Total $ cost of the materials needed to complete work order - Float
* Total Spot Purchase Cost – Total $ cost of the replacing item asset? – Float
* Total Miscellaneous Cost – Total $ cost of other needed to complete the work order - Float
* Description – description of the work performed, and further work needed – String - DROP
* Specialty Group – Specialty group hired to perform work on item asset before work order can be completed? – String (Categorical)