Big company logo

Analyzing Facilities Department Data

DSBA-6400 Internship Fall 202XX

DSBA Student Name

Doug Hague - Faculty Advisor
Name - Mentor

Table of Contents

Executive Summary

The internship is an opportunity to work with facilities department at Y company, analyzing data from their work management system. Machine learning can be applied to different areas of their business in order to improve operations. Some examples are predicting the cost and hours for a work request and how that could improve the scheduling of work, budgeting and providing better estimates to customers. Using natural language processing (NLP) to analyze the work request descriptions can improve how work requests are categorized into problem types and the speed at which they are routed to the appropriate shops to do the work. A big part of the work facilities department does involves installing, repairing and maintaining the equipment across the corporation. Analysis can be done to study the lifespan of equipment and see if there are any patterns of when equipment will break down. This can help avoid outages and highlight where additional preventive maintenance scheduling needs to be implemented.

The internship allowed me to apply different tools learned in the Data Science and Business Analytics program to different areas of facilities department operations and also get more experience working with real world data. The work management system generates a large volume of data where important insights can be lost if no one is there to analyze the data and turn it into knowledge. This can then help the facilities department make more informed data driven decisions going forward.

Introduction

Business Objectives

There were three main objectives that were the focus of this internship. The first was being able to provide better work request estimates in terms of the cost and hours they will take to complete using machine learning models. Having a data driven approach that will take into account different details about the work request when making estimates can improve budgeting, scheduling and customer expectations of when it will get completed.

The second objective is to improve and automate the categorization of work requests problems by analyzing the text descriptions of historical work request data. This improves the response time from when a work request is entered to when it gets routed to the shop performing the work. This would eliminate the manual intervention from the manager and allow them more time to focus on other important tasks.

The third objective is to analyze equipment data and find patterns where equipment has shorter than expected lifespans. This data can help facilities leadership make informed decisions on where to invest in new equipment, adding additional preventive maintenance to try to extend the life of the equipment and also budgeting when new equipment will likely need to be replaced.

Statement of Business Problem

One business problem facilities is faced with is having to respond to a large volume of work requests with limited resources. Time is a critical component because if there is a high priority problem on site it must be addressed quickly before it results in a larger problem which will increase the time and cost it takes to complete. Not all work requests are equal and there are different factors that contribute to the time estimate and cost to complete. By default, the time is estimated at 1.5 hours for most work requests. Supervisors or managers have the ability to change this, but this doesn't always happen. This is where machine learning can intervene to provide estimates by creating regression models on the previous historical work requests in order to make predictions on how long a work request should take to complete as well as the cost.

The business objective of categorizing problem types from customer entered work request descriptions, addresses a problem the facilities department currently faces where an application change no longer allows customers to enter problem types themselves. This was removed in order to minimize the number of fields required by the customer to fill out a work request. Knowing the problem type allows facilities to classify the type of work needed and route to the appropriate shop quicker. Speaking with a manager from the facilities department he estimated customers entered problem types correctly around 70% of the time and the manager could always override and change the classification. Data Science can be a solution to this problem by analyzing historical text descriptions and determining patterns that can classify the problem type with a high level of accuracy.

The facilities department is faced with maintaining different equipment across the site and determining when it should be replaced. Instead of making these decisions when equipment suddenly fails, being able to plan for it by analyzing equipment data can help facilities know when to budget for replacement. If certain brands and models of equipment perform worse over time than others, this can also be taken into consideration for what models to avoid when purchasing new equipment. Analysis can be done on the equipment and different fields such as equipment type can be analyzed to identify patterns of equipment being replaced after a certain period of time.

Relevant Literature

VS is the software where all the work request related information is stored. This includes the data dictionary to see the schema of work request related tables and each field's meaning, data type and values. There are also reports such as equipment lifetime and historical work requests that were referenced to understand the data and values represented in the tables.

Different types of machine learning models were researched for doing the analysis. Most notably, https://machinelearningmastery.com/ was one of the resources selected because it had

a number of great articles about different machine learning models and techniques such as XGBoost Regression, evaluation metrics and oversampling techniques.

Methods

Regression Analysis of Work Requests

Methods

Given the large volume of data, it was filtered going back a few years to 2015. The data has different information about the work request that was used in order to find different patterns when training the model and making predictions. The table below contains four different groups of work request data gathered for the analysis.

Type of Data	Features	Value
Date Information	● Requested Day ● Requested Month ● Requested Year ● Requested Day of the Week	Some parts of the year are busier than others in terms of the volume of work requests, therefore they may take more time to complete
Work Request Information	● Priority of the request ● Type of work ● Department\Sub department ● Account being charged	This captures valuable information about the type of work being performed and who is requesting the work.
Location Information	● Building ● Age of Building ● Room ● Room Type	Information about where a work request is located can help find patterns based on if a building is older or certain room types take longer to complete or more cost to complete than others.
Equipment Information	● Equipment Type ● Condition of equipment ● Model of equipment	Find patterns if work done on certain pieces of equipment takes longer to complete than others.

The data was cleaned so that all null values were removed. Most of the data is text so any duplicate categories containing misspelling or abbreviations were corrected. If categories could not be mapped to a value they were given a new category value of UNKNOWN or NOT_APPLICABLE if a field shouldn't have a value under certain conditions. Outliers represent extreme values that fall outside of the normal distribution of the rest of the data points, since these are sensitive to the regression results the outlier records were removed for the analysis.

To use machine learning models the data must be numerical, therefore the data was encoded using One Hot and Ordinal encoding techniques on the categorical fields. One Hot encoding creates a new column for every category in a field, so for example if you had a condition field

that had three categories: good, fair and poor, three columns would get created for each category that would be either 0 or 1. Ordinal encoding creates a numerical column and converts every category to a number based on an ordered relationship. For example, if there are three categories such as small, medium and large, you can ordinal encode the data so small is 0, medium is 1 and large is 2. The encoded data is then transformed into a sparse matrix completely numeric, free of null values and ready to be passed into the classification models for the machine learning.

A variety of regression models were researched such as RandomForest and Support Vector Regression, but the best performing model was **XGBoostRegressor** which is a popular ensemble boosting model. An ensemble boosting model is when multiple models are run to make predictions instead of relying on one model. With boosting a model it is run iteratively and during each run it tries to make corrections to the previous models errors with the goal to lower the amount of error during each iteration.

Two separate XGBoostRegressor models were run on different dependent variables, one model to predict the work request cost and the other to predict work request hours. The data is divided into a training set and a test set. The training set is used for model training and finds patterns in the data in order to make predictions. After the model is trained predictions are made on the test set. The test set represents data that has not been seen by the trained model, so it is essentially new data that will make a prediction using the patterns discovered by the model during training. The predictions are captured and compared to the actual values and the difference between the two is called the cost, which is the error between the two, with the goal being to reduce the cost as much as possible.

Barriers

When applying different regression models the main barrier was the time it took to train models with a large amount of data. A RandomForest model for example took over an hour to train on just a small sample of the data. This is what led to implementing an XGBoost model because it had both the advantage of being able to quickly train a model in a short period of time and produce lower error cost than the other models used.

Another time-related barrier is hyperparameter tuning. While it adds the benefit of potentially improving model performance, the cost is the time it takes to test different combinations of parameters. Some models have a large amount of parameters that can be changed but the more parameters that are tuned you are increasing the time it takes to train the parameter combinations.

NLP Classification of Problem Type

Methods

When work requests are entered into the system, customers have the ability to leave a detailed description about what the problem is. There is valuable information in this unstructured format

that can be extracted from the text and used to classify the problem. A method used for this analysis is Natural Language Processing, which is the process of analyzing a collection of words in order to derive meaningful information. The methods used in this section was to analyze the text and classify the work request under a problem type classification.

In the database, work request descriptions and problem types were pulled from data before March 20XX. After this date a change was implemented by the facilities department that prevented problem types from being entered by the customer. Therefore, data before this date represents the best data available for training the model. The next step was determining what problem types could be excluded. There are some problems types such as preventive maintenance that are automated. Since the main goal of this analysis is understanding customer generated text, problem types with auto generated text such as preventive maintenance were excluded.

There are over 300 problem types available and with data cleaning and combining similar problem types the number was reduced down to around 160. This is still a large number of categories to try and classify in a machine learning model and some problem types were only used once. Analysis on work request data for the top 20 problem types was used, because it provided the best sample size to train the model on.

To prepare the descriptions for text analysis a few common approaches were used to reduce the vocabulary. The first step was to convert everything to lowercase so there are not multiples of the same word in different case combinations, then any punctuation and numerical data was also removed so only text was remaining. The next step was to remove stop words which are words that occur frequently but do not add any relevance with respect to the classification and are just noise in the data. A technique called **Lemmatization** was used to take the word and reduce the base form of the word, as an example the lemmatized version of runs would return run. One additional step was to address the class imbalance, which is where the number of categories are not equal so you have a majority class with the most samples and one or more minority classes with lower amounts of samples. An oversampling technique called **Synthetic Minority Oversampling Technique (SMOTE)** was used that synthesizes samples from the minority classes to balance out with the number of examples in the majority class. The data is augmented in a way that makes it better than just simply duplicating examples to balance out the numbers.

The next step in text classification is creating a word embedding which is a representation of the text into a vector of numbers that can be used in machine learning classifiers. The vectors are representations of the words and the more similar the words are to each other the closer they are in the vector space. Each word in the cleaned text description field is broken up into a token, which is similar to a list of words. Common two-word and three-word combinations, known as bigram and trigrams respectively were also captured and added to the list. The list of words is then fed into a **Word2Vec** word embedding model and fitted using a neural network to measure the similarity words have to one another.

The Word2Vec model was then used in a deep learning classification model. The model is made up of different layers that each play a role in making a prediction on what problem to categorize the work request. The first layer is an embedding layer which are the values from the Word2Vec model which will act as weights. Two Bidirectional layers are then used and will handle the order of the words from left to right and vice versa. Lastly, at the end of the neural network two layers will be added to make the predictions using the probability. The model is trained on a sample of the data where it learns patterns in the text to try and classify the problem type correctly. The model then makes predictions on the unseen test data and then provides a probability for each problem type and the highest probability is chosen as what that work request gets classified as. Accuracy metrics, which will be discussed in the Results section can then be obtained to assess how well your model is performing.

Barriers

The main barrier to achieving better classification results is a combination of the large number of problem type classes and the imbalance between the majority and the minority classes. All relevant data was utilized in the analysis and unfortunately all new data coming in is not classified by problem type because of the new change they made to the system.

One possibility that has been discussed with facilities is merging similar problem types together. The existing problem type list is very detailed and one idea is that by combining problem types together you get the classes more balanced. An example of two problem types that exist are HVAC-TOOHOT and HVAC-TOOCOLD, since they are both related to room temperature they can be consolidated into a new ROOM-TEMP problem type.

Survival Analysis of Equipment

Methods

Equipment is a critical piece to the success of the facilities department and it must be properly maintained to extend the life of the equipment. An analysis was performed to see if there were any patterns linked to equipment and when it is replaced. **Survival Analysis** is a technique to analyze the time until an event occurs, which in this case is when a piece of equipment is placed out of service.

The first step is once again to gather the necessary data from the database. The facilities department has an equipment table for every piece of equipment both in and out of service. Different descriptive equipment fields such as the name, model, category, condition and status were captured in addition to the inservice date and outservice date if applicable. A binary event field was created called “is_end” that indicates if the equipment is out of service. A field capturing the number of years the equipment was in service was also created. The field is calculated by either subtracting the current year and inservice year if equipment is still in service, or subtracting the outservice year and inservice year if the equipment is out of service. A method known as **Kaplan-Meier** calculates the survival probability based on a duration and an event. The duration in this case is the number of years the equipment is in service and the

event is when the equipment is placed out of service which is in the “is_end” field. The Kaplan-Meier model is then fit to the data and an event table is generated with probability of survival over time.

The Kaplan-Meier model can be run at different levels of detail. It was first run on all the equipment data which range in service years from less than year to over 50. Then a more detailed analysis was done on specific categories of equipment that have a large sample size. The following table shows the top 3 equipment categories in terms of the number of records with outservice dates.

Equipment Category	Outservice Record Count
Equipment	1172
HVAC	344
Material Processing and Handling Equipment	232

Another level of detail was performed on each category capturing different equipment categories and examining how they differ in terms of survival probability. Below is an example of different equipment types analyzed for each equipment category.

Equipment Category	Equipment Type
Equipment	● Refrigeration ● Cooking ● Kitchen ● Dining
HVAC	● Chiller ● Boiler
Material Processing and Handling Equipment	● Utility Cart ● Commercial Unit

Barriers

A barrier when doing equipment analysis is that equipment can represent a wide variety of items from small to large with varying life expectancies. Certain items like a boiler could have a large life expectancy of 25 years but it is unlikely that a computer would have a similar life expectancy. Therefore identifying the type of equipment you are analyzing gives more perspective on whether a short life span is actually an issue when you see certain equipment being placed out of service earlier than others.

Another barrier is related to the availability and confidence of inservice dates. There were almost 20,000 pieces of equipment without in-service dates. Some equipment is older and they attempted to identify the correct inservice date to the best of their ability. They have a confidence level linked to the equipment based on the source of the date, so some inservice dates may not be as accurate which would lead to inaccurate survival probability results.

Lastly for some equipment the number of out of service records is not large, so one red flag this raises is if they haven't been recording the outservice dates for some equipment. This is why the analysis is done for the top number of features with a larger sample size. In speaking with the mentor he did say that some technicians do not always record the outservice date.

Mentor Role

The mentor and other members of the facilities department are important resources in understanding the work request process and knowing the limitations of the data. The internship involves working with a large amount of data and they have a wealth of institutional knowledge that is critical to help understand the meaning of certain terminology and how different processes work. The mentor was always available to answer any questions I had and work through any issues. The mentor will also play a role in how the work I have done gets implemented in the existing work request process workflow. We have discussed how the models can be used when work requests get created, where the model will take the values entered by the user as parameters and make a prediction for how much the work request will cost, the time it will take to complete and the problem type classification.

Results

Regression Analysis of Work Requests

The results of the regression analysis are measured in terms of the cost, which is the error between the predictions from the model and the actual value. There are different cost metrics available but the two chosen were **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** because they return the error in the same units of what is being predicted. The one difference between MAE and RMSE is that RMSE penalizes the results if there are larger error differences, so it will always be higher than MAE.

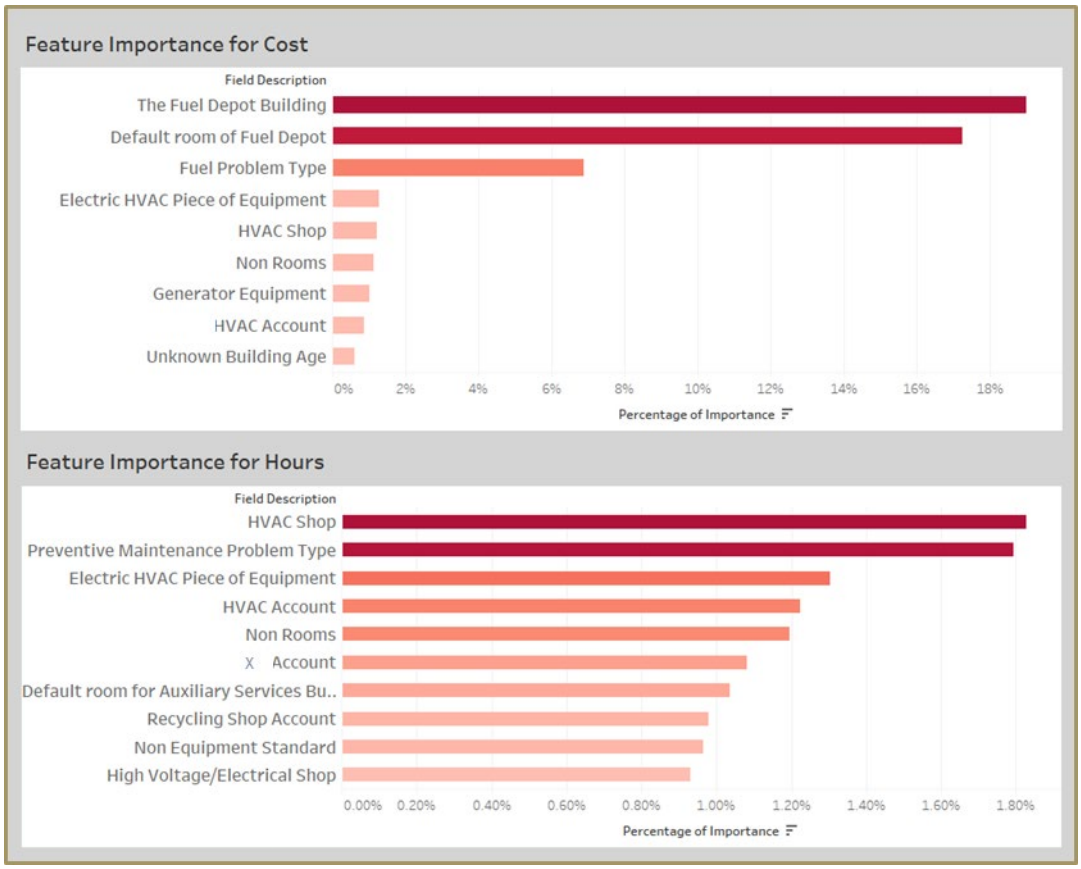
The following table shows the error for the regression analysis for cost and hours of work requests using both XGBoost models with and without cross validation. **Cross validation** is a way of running through different iterations of the model with different sample variations of the training set to train the model. The average of all the iterations is taken as the overall cost score. In the results below, they both generate similar error cost, which gives more reliability to the results. The prediction will yield an average error of $37 when predicting cost and .9 hours (54 minutes) when predicting hours.

Analysis Type	**Work Request Cost**		**Work Request Hours**	
XGBoost Regression	**MAE**	**RMSE**	**MAE**	**RMSE**
	$37.00	$57.24	.9 hours	1.28 hours
XGBoost Cross Validation (10 Folds)	**MAE**	**RMSE**	**MAE**	**RMSE**
	$37.10	$57.08	.9 hours	1.27 hours

The visualizations below show which features are most important in making predictions in order to provide explainability on how the model uses the features for predictions. Since there are a large number of features used to train the model, only the top features are shown.

The first visualization shows the importance for work request cost. The top three fields are related to Fuel operations and represent 43% of the importance out of all the features. Other noticeable features in the top 10 are around HVAC, where the HVAC shop, equipment and accounts are shown for the top importance.

The second visualization shows the importance for work request hours. The importance is distributed across more fields, where the top feature only accounts for 1.83% of the importance. There are similarities in fields between predicting cost and hours, such as HVAC shops, equipment and accounts, but there are also new features in the top such as the Preventive Maintenance problem type and different accounts

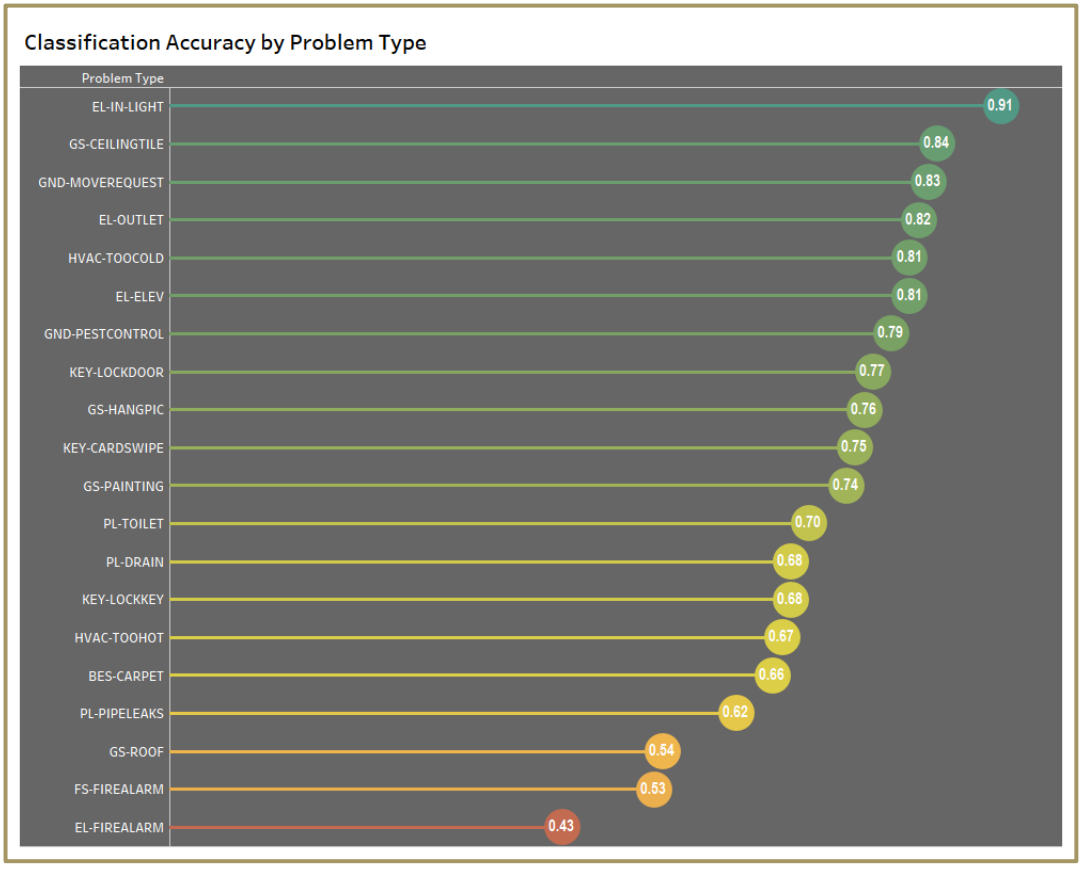


NLP Classification of Problem Type

The model was trained on all available data for the top 20 problem types before March 3, 20XX, when facilities made the application change resulting in them no longer recording problem types. The classification analysis resulted in an overall accuracy of 75%. The following

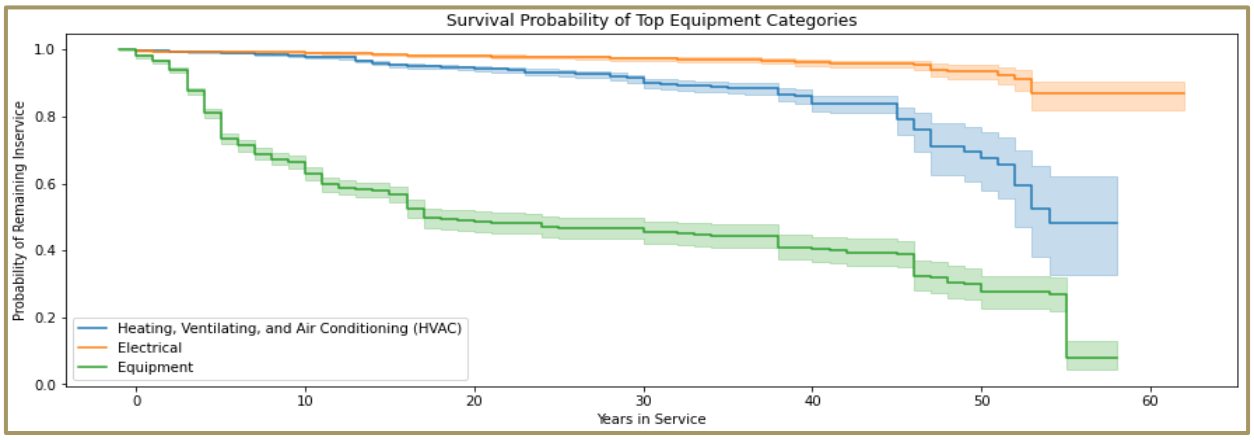
visualization shows the breakdown of accuracy for each problem type. Overall the best accuracy was 91%, made on EL-IN-LIGHT which indicate electrical issues related to indoor lighting and the lowest was 43% for EL-FIREALARM which indicates work on fire alarm electrical. Usually the categories with better results had the highest amount of sample data available.

One issue encountered with this task was dealing with the large list of problem types they have in their work request data. We have discussed coming up with a way to have a more general list of problem types that will be used in the classification and keeping more detailed problem types in a separate field. Additional work is needed to get more data to train on in order to get better classification results. A recommendation that came out of this analysis is to work with the facilities department to consolidate the problem types into smaller groups, which will allow for more training data, because more samples will be combined into a more manageable number of problem categories. This will hopefully improve the classification accuracy for problem types that currently have low accuracy results. Some examples highlighted in the visualization below show HVAC-TOOCOLD has a .81 accuracy, but HVAC-TOOHOT is much lower at .67. If they were combined the hope is that it can improve the overall accuracy. Another example are the bottom two problem types related to fire alarms, if facilities agrees that these can get merged under one problem type, hopefully the combined samples will result in higher accuracy.

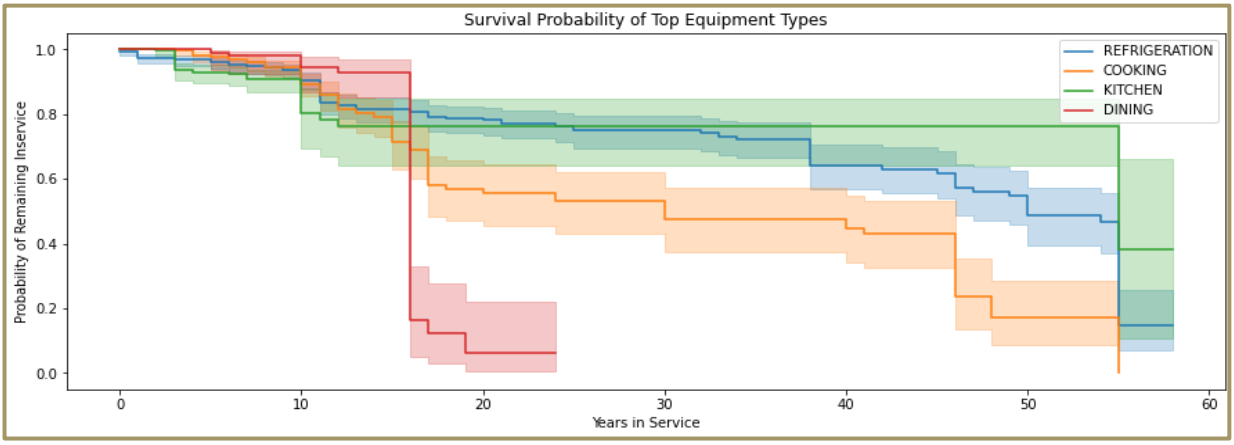


Survival Analysis of Equipment

The first results show the survival probability categorized by equipment category. The top three categories in terms of the highest number of equipment that has been placed out of service was chosen. The figure below shows that the equipment category (green) has the sharpest decline in survival probability over time and that the electrical equipment (orange) has the best survival probability over time.

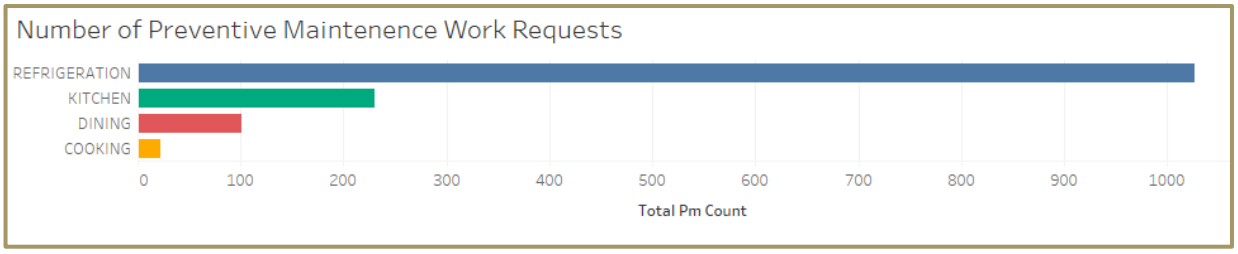


Since the equipment category had the sharpest drop off of all of the examples examined, the top equipment types for this category were looked into further. The following chart shows the top Equipment category types and the results show that Dining equipment (red) has the sharpest drop in survival probability over time at around 17 years.



Of the 124 records that are classified as Dining equipment, 24 have been placed out of service. More analysis was done examining the number of preventive maintenance (PM) work requests which are work requests that are routinely scheduled after a certain period of time to maintain equipment. The graph below shows that of the four types of equipment in the example above, both Dining and Cooking are significantly lower than Refrigeration and Kitchen in terms of the number of PM work requests linked to that type of equipment. Dining and Cooking also have the two lowest survival probabilities over time. It appears that there could be some correlation

with PM work requests extending the years of service of these pieces of equipment. A more long-term study would need to be implemented to see if adding more PM's results in extended life.



Discussions and Conclusions

Overall Business Impact

Overall the analysis done can improve upon different areas such as estimating time, allocation of resource hours, budgeting cost and staff and where improvements can be made on equipment. Instead of defaulting work requests to an arbitrary number of hours and cost and relying on a manager to change it manually later, data driven estimates can be predicted from models that were developed during the internship. This can provide more realistic expectations to the customer about when their request will get completed and allow for better resource management by knowing how many hours a technician should be allocated to a certain task.

Predicting the problem types of work requests is a solution to a recent business problem, because customers are no longer allowed to select problem types when entering work requests. The model can automatically assign problem types from the description, which will result in less manual intervention and the ability to route work to appropriate shops quicker.

Analyzing the lifespan of different equipment has a more long-term impact because facilities leadership can identify patterns where certain types of equipment get replaced quicker than anticipated. There are also reactive measures that can be taken from the analysis such as implementing more PM procedures on at-risk equipment to try and extend the life.

Recommendations and Next Steps

The next phase is working with the VS developers on implementing the machine learning models that were created during the internship. The work request cost and time prediction model will be integrated into the VS work request creation form where the user enters data and will be used as parameters in making predictions.

Based on what was found from the problem type analysis, additional work will be needed to merge existing problem types into more general groups to improve text classification results. A recommendation is having different fields that represent different levels of detail for problem types. A more general field of problem types would contain a small list of problems that will be

used for classification and then as technicians complete work they can fill out more specific problem details in different fields further down the process. Once this is complete the text classification model can be retrained and then implemented into the VS work request creation process.

The findings of the survival analysis show a sharp drop in survival for Dining equipment and there were a low number of Preventive Maintenance (PM) work requests while longer lasting equipment had more PM's. A recommendation is that a long-term study be done that would implement more PM's for Dining equipment and over time see if that extends the lifespan of the equipment. The one challenge in doing the equipment analysis was the availability of data because inservice and outservice fields were not always filled out and could not be used in the analysis. Facilities would need to make this a required part of their business process going forward in order to have more detailed analysis done on equipment.

Appendix

Internship Related to DSBA Program Courses

The internship covered many different areas of Data Science that were taught in the DSBA courses I have taken. The first step of the process is getting the data and I was able to use the SQL knowledge I have gained in the Big Data Design, Storage and Provenance course to query the data needed for analysis.

In the Applied Machine Learning course I learned about different data cleaning techniques that I was able to use to clean the data and make it numerical for use in machine learning models. I learned about how different encoding techniques were applied and a Pipeline was used to chain the encoding techniques and feed that into the machine learning models. Those techniques were all used in my internship. In both Applied Machine Learning and Business Analytics courses I learned about the different machine learning models that are used in Data Science such as Word2Vec and the pros and cons and when it is applicable to use one over the other. This knowledge I learned in these classes helped me when researching and choosing the model for each task.

An important part of data science is storytelling and how to transform the machine learning results to something more visual that people less technical can easily understand. Visualizations are a key part of storytelling and the Visual Analytics course taught me about the different types of visualizations, design techniques and how to create dashboards to get your point across to the viewer. Another part of storytelling is that most DSBA courses I have taken have a project component to them. The presentations we give help us improve our storytelling skills and this is an important part of Data Science. I feel the courses I have taken over the years helped prepare me better for future presentations and being able to discuss findings with facilities leadership. Overall this internship was a great opportunity to apply what I learned in many different ways.

Internship Experience

Overall the internship experience was very rewarding because I was able to apply the skills I learned to solve real world problems. Having this experience and knowing how to develop a plan of action to build solutions and solve problems is a very valuable tool to have when moving on from Student to Data Scientist. The internship allowed me to learn new techniques through research, such as how to deal with class imbalance for classification problems or different hyperparameter tuning methods. It is also great to know that the work I have done during this internship can help make an impact in the university and improve certain operations within the facilities department.