AN INTELLIGENT DATA-DRIVEN ANALYTICS SYSTEM FOR OPERATION MANAGEMENT, BUDGETING, AND RESOURCE ALLOCATION USING MACHINE LEARNING AND DATA ANALYTICS

Dele Fei[1] and Yu Sun[2]

[1]St. Margaret's Episcopal School, 31641 La Novia Avenue,
San Juan Capistrano, CA 92675
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

This is a data science project for a manufacturing company in China [1]. The task was to forecast the likelihood that each product would need repair or service by a technician in order to forecast how often the products would need to be serviced after they were installed. That forecast could then be used to estimate the correct price for selling a product warranty [2]. The underlying forecast model in the R Programming language for all of the companies products is established. In addition, an interactive web app using R Shiny is developed so the business could see the forecast and recommended warranty price for each of their products and customer types [3]. The user can select a product and customer type and input the number of products and the web app displays charts and tables that show the probability of the product needing service over time, the forecasted costs of service, along with potential income and the recommended warranty price.

KEYWORDS

Operation Management, Machine Learning, Data Mining.

1. INTRODUCTION

This research was based on a manufacturing company's service department. The company's name is FastLink China. The products it produces are mostly industrial doors and dock levelers [4]. It is the top 1 in this criteria of business in China. Since it is still a developing company, there are hopes to make it better in minor parts. The goal for this project is to analyze the best price to maximize the profit in the service department. It is imperative to the company simply because the service department has the highest profit rate compared to other departments. However, the problem is that company owners have no idea how to set a price for extended warranty to help them gain the profit. Indeed, the research is based on the hope that this problem could be solved through data science skills. This can lead to a much larger profit in the company when they know the exact cost for the warranty and then determine profit based on different types of customers and industrial doors. Besides, it is a precious opportunity for me to learn about data application on businesses and to get familiar with how it runs and the way it works.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 45-56, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111904

There are a variety of tools or systems like fire base that have been used as a means for users to analyze their data for their personal use or business use. However, these existing tools are not that useful to me. Their implication and usage are too fundamental for the case since this is a research by setting up an analysis for a unique kind of data for a unique type of business [5]. It is not common that these existing tools seem limited. If the final result is provided by these tools, the accuracy will be doubtful and might have a huge influence since it is provided to a company that is related to money [6]. In this case, a more customized and sophisticated analytic tool is needed to provide a reliable and credible result to the company.

On the other hand, there is machine learning that is obviously workable in this situation. However, it is time consuming and too technical for a high school student to perfect the result also due to limitation of resources. Moreover, it is tough to explain the logic behind it to the people in the business company. Therefore, finding a better tool than normal, but simpler than machine learning is the goal [7].

Our goal is to forecast the price of warranties that will benefit the company. To this end, a survival curve is used inspired by insurance companies. Speaking of tools, the front end and back end both exists for the users. Front end is more HTML coding with apex charts and graphically displacement. For the back-end, tidy-verse and r are used for basic data cleaning. From ggplots and survival curve, the predicted percentage of breakdown for each month of each type of productcan be accessed [9]. Then, with basic calculations and taking the mean value of costs for appointments, an reliable outcome of recommended prices of warranty can be produced.

To the end of proving the result, admittedly, the most common ways will be to use train and test to see which has the best prediction. Also, r squared is a value that is usually considered by data scientists [8]. However, the situation is different from others since data that can be considered as correct to compare with the warranties does not exist. Moreover, there isn't enough data for us to train and split in some situations since some types of business and products only have a small amount of service history after group buys. Besides, the plot shows a strong curve just by plotting it through ggplots. Indeed, the research uses the approach to predict it through user surveys because the opportunity to let the manager class in the company determine if this data is applicable in real life situations or not exists. This is useful because all the people who took the survey will be familiar with all the products, companies, and potential users, which is the service department in this case.

Through looking at the official definition of user survey, a survey with 10 questions each with a rating from one to five is designed. After adding the rating for positive questions like "I think that I would like to use this system frequently", and minusing the rating for negative questions like "I found the system unnecessarily complex," the total score is doubled to normalize the total score like the official website asks. Indeed, getting the result over 68, which is the number to determine if the model is useful, proves that this development is effective.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges during the experiment and designing the sample; Section 3 focuses on the details of solutions corresponding to the challenges that was mentioned in Section 2; Section 4 presents the relevant details about the experiment, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Identifying the Problem and Approach

The most important challenge in this problem is identifying the problem and finding an approach to it. Beginning with a single conversation, the company owner complained about how it is tough for them to set a proper price for the extended warranty. Without it, he is concerned about how to plan for his companies' future since he does not have a predicted revenue. Instead of having the concern go in one ear and out the other, intuitively realize that data science could solve this problem, curiosity drives to investigate in how the price of the extended warranty with existing data can be forecasted [10]. Then based on the existing data, a unique approach emerges. After communication, the date of past services, cost and income of each service, and when the door starts to function exists in the database. Even though the realization of that number of dates is a vital variable to the prediction helps to get on the right path, going through linear models, logistic models, machine learning, and neutral networks, the correct approach is still ambiguous to this question. While the question lingers in mind, a collaborator who works in a data science company help on deciding a survival curve is usually used for the predictions for insurance companies that has the input of time elapsed for each product. Then, the connection between the average cost and benefit for each product will make the final step toward the final result.

2.2. Setting Price

One challenge in the problem is how to set a price for the extended warranty that is acceptable to the customers and profitable for the company. The company expresses their concern that they are only making guesses for the price that should be set for the warranty because they have no idea on the probability of the chance the specific type of product is going to break down. This problem is vital to the company's income because competing companies set a cheap price for their product, but an expensive price on their extended warranty [11]. To maintain the market, the company needs to set a price that is able to comfort the customers. In addition, they have never calculated the mean value of cost of labor and parts for fixing. Indeed, this model helps to solve this problem by investigating the number of breakdowns in an order for a period of time to get the percentage and uses the average cost to determine a price for the future orders.

2.3. Data

Initially, jetlag and distance between China and the United States forms a communication issue that makes the process of getting the data hard. Eventually, repeating the process of waiting and asking for more data, final result is slowly approached as the research develops. However, challenges come with opportunities. Since the workers of the company have not practiced standardized training on entering the data, not only the format of data is usually a mess to deal with, the emptiness also becomes a hot potato of the research. Needing to select a boundaries of data that eliminates the outliers and empty data without too much influence on the result is important. Also, there is not much data from the company so it is imperative to try to keep the amount of data [12]. Through identifying and excluding the empty data, the data for the initial prediction is cleaned. However, there are still some services that produce a negative profit for the company, which does not make sense. In this case, the company itself needs to find the correct input in order to solve this problem successfully.

3. SOLUTION

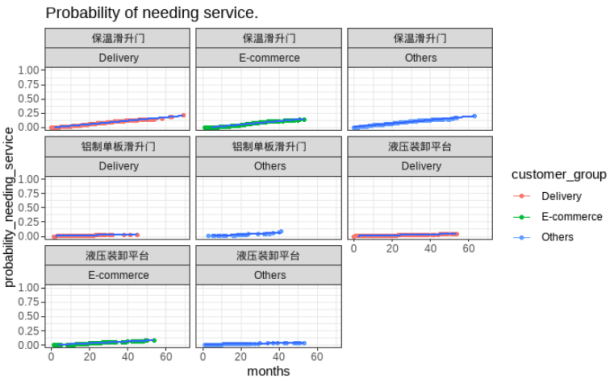

Figure 1. Probability of needing service

The scope of this project is to address the problems of … To break down the challenge into smaller programmable questions, these questions were generated:

What is the final outcome/indices that I'm trying to compute? What is necessary for me to compute such indices?

1. maintenance: (% of product need maintenance, and cost per maintenance [whether through average or median, and why did you choose it])
2. replacement: same as above

In general, the recommended price of extended warranty that can make the company predict the probability of breakdowns is the goal of this research. With the data of the door starting functioning date, needing service date, a survival curve is used like other insurance companies mostly use. Since the goal is to output the recommended warranty price, getting the center value of past historical services to know what the predicting cost and profit is required by combining two data. Eventually, the outliers are eliminated and just take the mean since the data is not heavily skewed. Indeed, the profit percentage is an input for the user for what they are looking for.

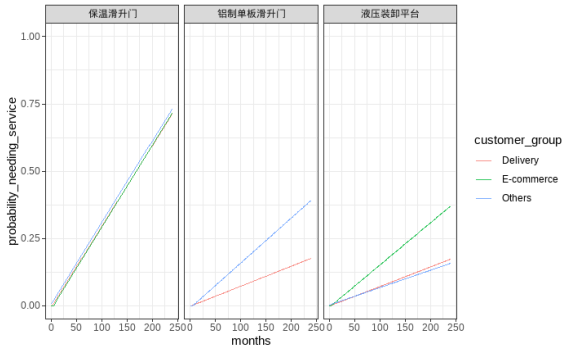

Figure 2. Breakdown percentage

Going into more detail step by step, graph of breakdown percentage is established using existing data to predict the trend. The x-axis represents the months, and the y-axis represents the percentage of this door needing service. The graph is divided into three products. Each of them is showing their major type of customer, which is normally the one that has enough data to predict

the trend. Since the business has the most customer groups in delivery and E-commerce, both types have the sufficient amount of data to make predictions. From the figure 2, the data has a strong correlation. It is great for future predictions.

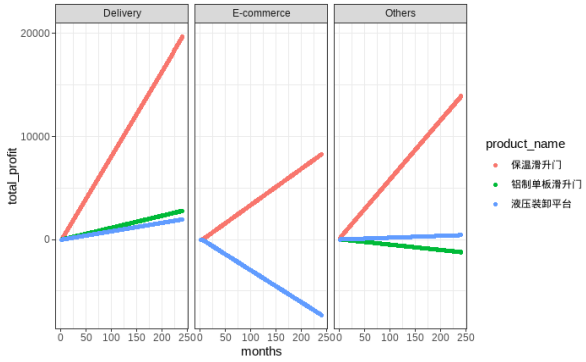

Figure 3. Profit

After getting the average cost and profit from the past data, these data with the percentage earlier are implemented to see the total profit that the business could gain after months for specific products in different businesses. This would be useful for the business since they could have a visualized understanding of its profit in the service department, which can be used for adjusting their price or percentage of profit based on their past data. It is obvious that there is some error in the existing data since there are some products that are bringing negative profits. This error might be due to the incorrect input from workers or the incorrect price they set for the customers by the business.

product_name <chr>	customer_group <chr>	n_total_products <dbl>	appointments_per_year <dbl>
保温滑升门	Delivery	100	3.6003095
保温滑升门	E-commerce	100	3.6674700
保温滑升门	Others	100	3.6426980
铝制单板滑升门	Delivery	100	0.8922874
铝制单板滑升门	Others	100	2.0091848
液压装卸平台	Delivery	100	0.8735344
液压装卸平台	E-commerce	100	1.8829967
液压装卸平台	Others	100	0.7751315

Figure 4. Appointments of each group

break_even_warranty_price <dbl>	recommended_warranty_price <dbl>	warranty_profit <dbl>
1301.2025	1951.8038	650.6013
1151.5917	1727.3876	575.7959
1375.2239	2062.8359	687.6120
329.9657	494.9486	164.9829
528.5626	792.8439	264.2813
506.3778	759.5667	253.1889
1060.6087	1590.9131	530.3044
536.5768	804.8653	268.2884

Figure 5. Prices and profit

Another feature the research produces for the company is a table that allows users to input the number of products as n_total_products and the percentage of profit they want to make. Since the model of percentage of needing services and the average cost of each appointment for different products both existed, this table is able to produce a recommended warranty price taking the percentage of profit the user is hoping to make. Also, the last graph of the businesses existing profit can assist in their decision to adjust the percentage of profits based on different products.

The table is also divided into three types of products and different types of customers. This is useful to help the company to set the warranty price.

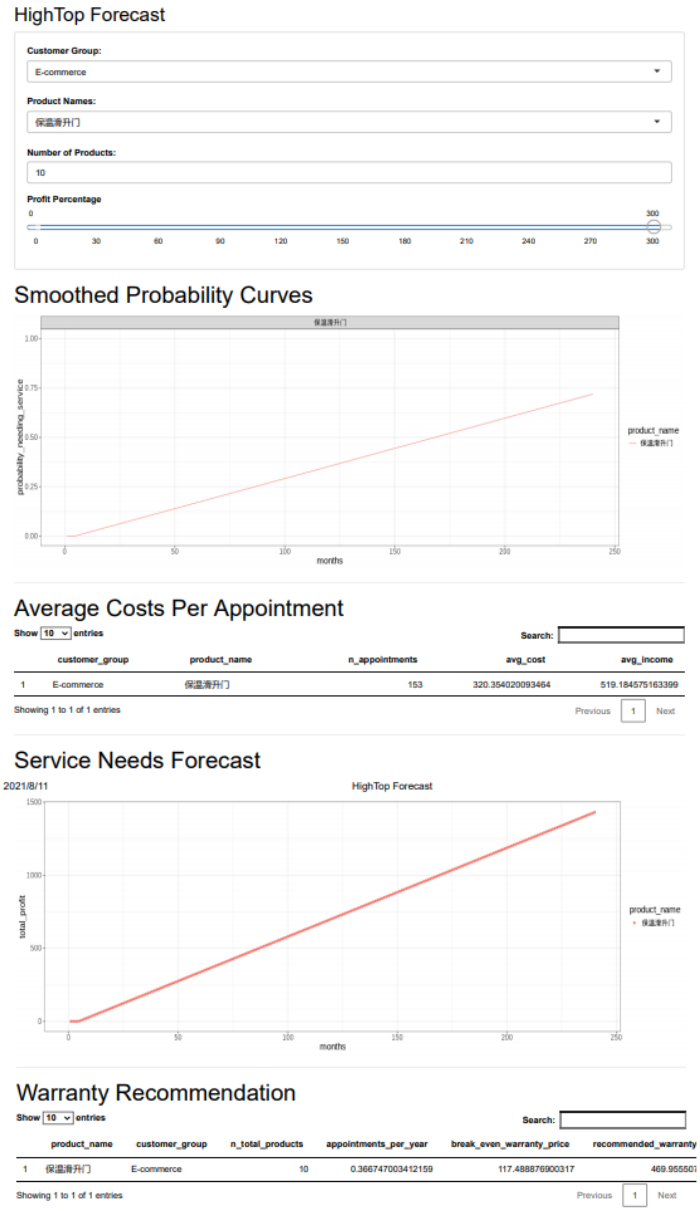

Figure 6. Warranty price

Besides pure coding, front-ends exists that directly help the users or business to understand the information they want. The input part allows business to adjust the customers' types, the products they are selling, the total number of products included in the warranty, and the percentage of profit. Then, the existing data will demonstrate a table and a chart. The chart displays the probability of needing services specifically that product and the type of customer. The table represents the average cost, profit, income of the specific product. Finally, it displays a graph of the probability of the breakdown, which is a prediction based on all the service history of this product. Last but not least, a warranty price recommendation is there for the user in the bottom table.

The choice of establishing a front-end is to better demonstrate data and encourage all the services workers to utilize it since it will be helpful eventually. If all of them are coding, the valuable information can only be accessible to a small number of workers, which decreases the efficiency of working.

In conclusion, this table fits the goal of displaying all the information that is helpful for the company to know their past data and the recommended future price for their future planning.

```
format_data <- function(){

  # this function takes the loaded data frames and cleans and prepares it so there is one row per product
  cat("Formatting data.", fill = T)
  service_data <- order_information %>%
    filter(product_name %in% c("保温滑升门",
                               "液压踏知平台",
                               "铝制单板滑升门")) %>%
    mutate(month_install = as.Date(paste(month(install_date), '1', year(install_date), sep = "/"), format = "%m/%d/%Y")) %>%
    group_by(customer_type, project_name, customer, product_name, month_install, warranty_length_zh) %>%
    summarize(
      install_date = min(install_date),
      product_quantity = sum(product_quantity)
      ) %>%
    select(-month_install) %>%
    left_join(
      google_translations %>%
        rename(warranty_length_zh = chinese,
               warranty = numeric) %>%
        select(-google_translate)
      ) %>%
    mutate(warranty_end_date = install_date %m+% years(warranty)) %>%
    uncount(product_quantity) %>%
    group_by(customer_type, project_name, customer, product_name, install_date) %>%
    mutate(item_number = 1:n()) %>%
    left_join(
      service_log %>%
        select(-project_zone, -customer_name, -city) %>%
        group_by(project_name, product_name) %>%
        mutate(item_number = 1:n())
      ) %>%
    left_join(service_type) %>%
    mutate(product_type = case_when(
      product_name == '保温滑升门' ~ "insulated door",
      product_name == "液压踏知平台" ~ "dock lever",
      product_name == "铝制单板滑升门" ~ "aluminum door"
      )) %>%
    mutate(status = ifelse(!is.na(service_date), 1, 0),
           days = ifelse(
             !is.na(service_date),
             as.numeric(as.Date(service_date) - as.Date(install_date), units = "days"),
             as.numeric(as.Date("2021-04-22") - as.Date(install_date), units = "days")
             )) %>%
    mutate(man_made_damage = ifelse(is.na(man_made_damage), 0, man_made_damage))  %>%
    mutate(out_of_warranty = case_when(
      service_date > warranty_end_date & man_made_damage == 0 ~ 1,
      service_date <= warranty_end_date & man_made_damage == 0 ~ 0,
      is.na(service_date) & man_made_damage == 0 ~ 0
      ))

  cat("Adding customer groups.", fill = T)
  customer_groups <- create_customer_groups(service_data, cutoff = 100)
  customer_groups <- customer_groups$customer_groups %>%
    select(product_name, customer_type, customer_group)

  service_data <- service_data %>%
    inner_join(customer_groups) %>%
    distinct

  cat("Returning data.", fill = T)
  return(service_data)
}
```

Figure 7. Code of function

First, the implementation with data acquisition is initialized. To do this in R, survival curve, ggplot, tidyverse packages are needed [13]. The data type is in the csv format where 21 columns are presented and in total there are 5296 rows of data. To format this data R based fundamental coding and tidyverse is used, where the data is formatted by grouping by three basic products and formatting the date types according to the data given. Through calculation, having a dataset for the number of dates the product lasted for the survival curve later on can be achieved. Also, any man-made damage or existing warranty services was taken out since they do not have a proper cost for later calculations.

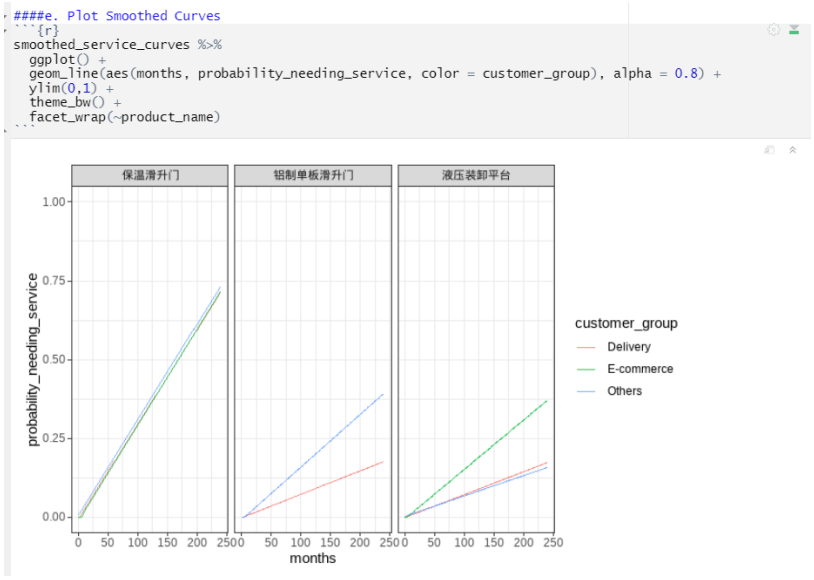

Figure 8. Code of plot smoothed curves

After getting and cleaning the data, the next step is to do some plotting the data to find the trend. a survival curve is used to generate a plot (figure 8) of probability of breakdowns during months according to different types of products and customers. Figure 8 is the result after ggplotting it and smoothing the curve. It will be later prepared with the average cost of services to calculate the recommended price of warranties.

```r
# forecast service costs
forecast_service_needs <- function(smoothed_curves,
                                   number_products = 100,
                                   financial_data){
  smoothed_curves %>%
    mutate_at(vars(product_name, customer_group), trimws) %>%
    mutate(total_products_needing_service = probability_needing_service * number_products) %>%
    mutate(new_products_needing_service = lead(total_products_needing_service) -
total_products_needing_service) %>%
    left_join(financial_data %>%
                mutate_at(vars(product_name, customer_group), trimws) %>%
                select(customer_group, product_name, avg_cost, avg_income)) %>%
    mutate(total_cost = total_products_needing_service * avg_cost,
           total_income = total_products_needing_service * avg_income) %>%
    select(-avg_cost, -avg_income) %>%
    mutate(total_profit = total_income - total_cost)
```

Figure 9. Code of forecast service costs

Then, this is a function that produces a data table that has values of predicting services needed for a hundred products, the percentage the average cost, income, and profit of the product combining with the outcome of the smoothen survival curve just produced.

```r
# calc recommended warranty

calc_recommended_warranty <- function(service_data = service_needs,
                                      financial_data = avg_financial_data,
                                      n_products = 100,
                                      profit_percentage = 1){

  service_data %>%
    mutate(year = ceiling(months / 12)) %>%
    group_by(product_name, customer_group, year) %>%
    summarize(total_probability_needing_service = max(probability_needing_service)) %>%
    mutate(n_total_products = n_products,
           projected_total_appointments = n_products * total_probability_needing_service) %>%
    mutate(appointments_per_year = lead(projected_total_appointments) - projected_total_appointments) %>%
    na.omit %>%
    inner_join(financial_data %>% select(customer_group, product_name, avg_cost)) %>%
    mutate(total_avg_running_cost = projected_total_appointments * avg_cost,
           avg_yearly_cost = appointments_per_year * avg_cost) %>%
    ungroup %>%
    select(product_name, customer_group, n_total_products, appointments_per_year, avg_yearly_cost) %>%
    distinct(product_name, customer_group, .keep_all = T) %>%
    rename(break_even_warranty_price = avg_yearly_cost) %>%
    mutate(recommended_warranty_price = break_even_warranty_price + (break_even_warranty_price *
profit_percentage),
           warranty_profit = recommended_warranty_price - break_even_warranty_price)


}
```

Figure 10. Code of recommended warranty

Eventually, a recommended warranty based on the profit percentage as an input is calculated. Based on the previously calculated cost, this code to produce a table of probability of breakdown and the trend of cost, income, profit by the different products and different customer groups is used.

4. EXPERIMENT

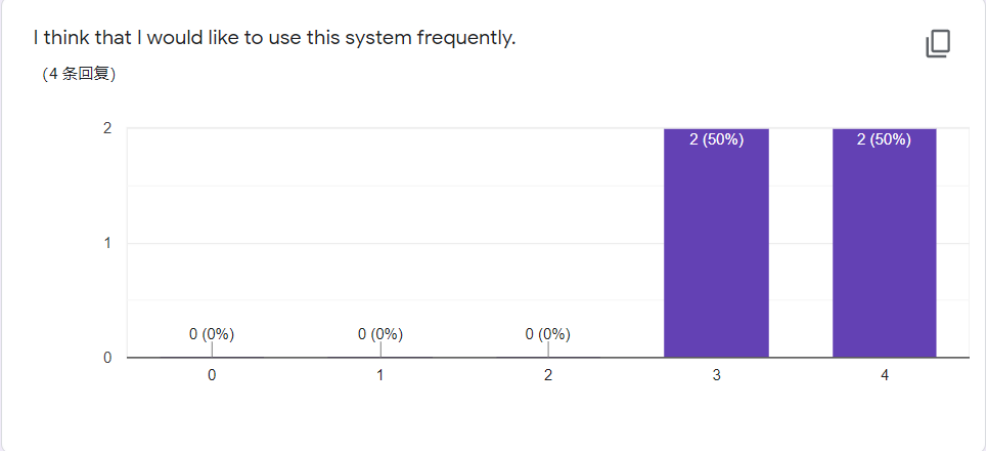

Figure 11. Survey result 1

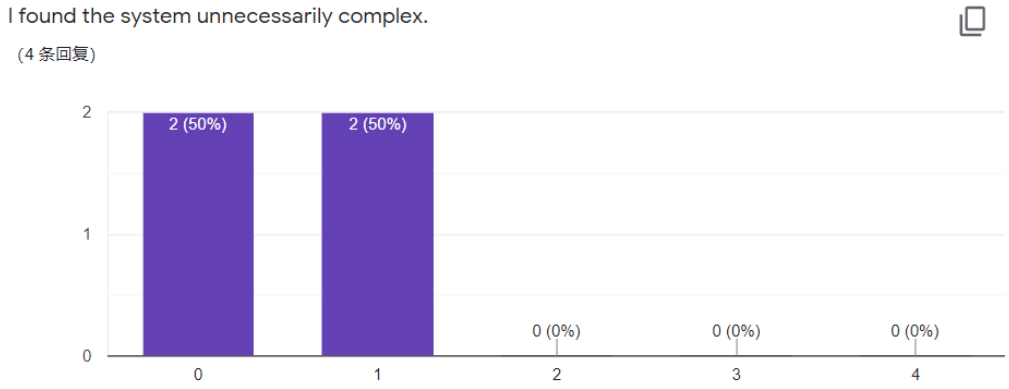

Figure 12. Survey result 2

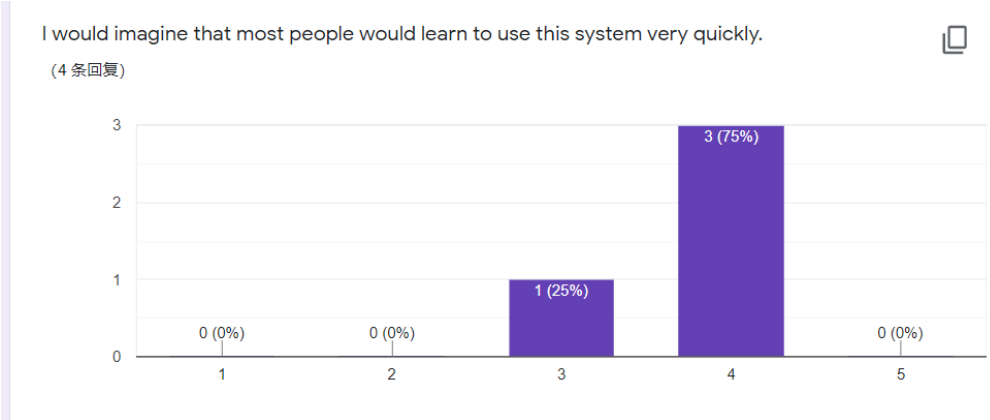

Figure 13. Survey result 3

The percentage of breakdown for each product is calculated. Including product types and business types as categorical variables, group by is used to filter out the history data that is specific to the product type and business types. After, the dates between the date the product are calculated that needs service to the date the product starts functions. Applying these variables into the survival model, which is a model that is usually used to calculate warranty for insurance companies, the regression line can be find that predicts the increase of probability in months. One interesting observation from the result is illustrated in figure 12, where the users rated the system based on its complexity. Since the overall system only has few pages, this question received a low score. This is output as how many services an order may need in x number of months in the table by multiplying the probability and the number of products in this order.

From this output, a test dataset can be use to find out if this data is correctly predicting in the real world situation.

5. RELATED WORK

A warranty forecasting model based on piecewise statistical distributions and stochastic simulation under the circumstance of having a large amount of data for the services already. Giving a interesting methodology to a similar question [14].

Another interesting domain in the field is the xxx presented by xxx in 2000. In general this paper forecasts the number of warranties through two phases. In phase I, they find upper and lower bounds of the warranty claim rates [15]. In phase II, they forecast for the recently launched product through the bounds in Phase I with a model built with the NHPP (non-homogeneous Poisson process) and the constrained maximum likelihood estimation.

This paper presents a forecasting method to predict a service system's expected number of through observed data which is used to calibrate a Generalized Renewal Processes (GRP) model [16]. It goes into detail of how the production in different months may impact the possibility of failures in the cars.

6. CONCLUSIONS

Warranty pricing is important to many businesses in different industry. For example, cars and headphones. Since the demand of having a warranty is huge, addressing the problem of how to price the warranty arises. Initially, initialized from a problem in a conversation, the research propose to bring a solution to the concern. Looking at the data from the department, an appropriate approach is identified after a conversation with the collaborator. A variety of solutions or proposals to solve this problem come up during the process. However, the survival curve is chosen because it fits the demand the best and it has been used in warranty companies. Later, the connection between the price and curve to produce the solution is last step needed.

Beginning with cleaning data, identifying different types of products, the subtraction between dates, and what data is man-made damage is needed since these data should not count. Before forecasting a recommended warranty price, the cost of services of products during a period is predicted. Using survival curves and ggplots, the prediction of the probability of breakdown is found. Combining the curve and costs, a recommended price of warranty to the users can be given eventually.

After all, user surveys as the experiment are used since not only the data has a strong correlation already in the plot, the existing warranty price does not give a good measurement on if the price is correct since it is pure guessing because the existence of the project is to help the company determine a correct price. Indeed, the experiment shows that the solution is effective and will help the company to envision its future success. Then, this is a successful data science project.

Current limitation is that the data sample is not enough. Since the service history system just came out two years ago for the company. Two year’s data cannot be adequate enough to predict all the trends after dividing them into different groups by business types and product types. Then the practicability of this model is being doubted.

Admittedly the data sample is too small, it can be solved very soon since FastLink is a fast growing company that has an enormous amount of data coming daily. Indeed, this new data can be used to optimize this model to predict better.

The system’s usability can be tested in the future. Through comparison between the date products actually breaks down in the future to see if the model works. If there is new data, the train dataset can be adjusted to see which will have the least error with the new incoming data. It will be a repeating process of iteration and optimization.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Jonathan Fei and Lina Tang for this opportunity to work on their data. Also, I'm deeply indebted to I would also like to extend my deepest gratitude to Beau Walker who helps me during the process of writing the code

REFERENCES

[1] Saltz, Jeffrey, and Kevin Crowston. "Comparing data science project management methodologies via a controlled experiment." (2017).
[2] Blischke, Wallace, ed. Warranty cost analysis. CRC Press, 2019.
[3] Newman, William M., and Michael G. Lamming. Interactive system design. Reading: Addison-Wesley, 1995.
[4] Hahn, Norbert, and R. Holzhauer. "Comparing dock levelers." Plant Engineering 50.11 (1996): 64-67.
[5] Johnson-Laird, Philip N., and Joanna Tagart. "How implication is understood." The American Journal of Psychology 82.3 (1969): 367-373.
[6] Freese, Frank. "Testing accuracy." Forest Science 6.2 (1960): 139-45.
[7] Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.
[8] Cameron, A. Colin, and Frank AG Windmeijer. "An R-squared measure of goodness of fit for some common nonlinear regression models." Journal of econometrics 77.2 (1997): 329-342.
[9] Wickham, Hadley, and Maintainer Hadley Wickham. "The ggplot package." Google Scholar. http://ftp. uni-bayreuth. de/math/statlib/R/CRAN/doc/packages/ggplot. pdf (2007).
[10] West, Kenneth D. "Forecast evaluation." Handbook of economic forecasting 1 (2006): 99-134.
[11] Baudrillard, Jean. The vital illusion. Columbia University Press, 2000.
[12] Lingis, Alphonso. The imperative. Indiana University Press, 1998.
[13] Wickham, Hadley. "The tidyverse." R package ver 1.1 (2017): 836.
[14] Kleyner, Andre, and Peter Sandborn. "A warranty forecasting model based on piecewise statistical distributions and stochastic simulation." Reliability Engineering & System Safety 88.3 (2005): 207-214.
[15] Wu, Shaomin, and Artur Akbarov. "Forecasting warranty claims for recently launched products." Reliability Engineering & System Safety 106 (2012): 160-164.
[16] Koutsellis, Themistoklis, et al. "Warranty forecasting of repairable systems for different production patterns." SAE International Journal of Materials and Manufacturing 10.3 (2017): 264-273.