Data Science - GA

Diana F. Camacho

11/9/17

**Project Design Writeup and Approval**

Follow this as a guide to completing the project design writeup. The questions for each section are merely there to suggest what the baseline should cover; be sure to use detail as it will make the project much easier to approach as the class moves on.

**Project Problem and Hypothesis**

* **What's the project about?**

My project is about mishandled bags in the airline industry. I cannot say the name or the airline or the name of the airport station, therefore I will refer to the name of the airline as PMT and the airport station as HUB1. Just know that it’s a well-known airline and a big international airport somewhere in the US.

* **What problem are you solving?**

The problem I am solving is identify opportunities to reduce the number of mishandled bags. The key metric PMT uses is called, “Mishandled Bag Ratio” or MBR, it’s a mishandled bag per every thousand customers.

* **Where does this seem to reside as a machine learning problem?**

Because I have collected almost two-year data that can be used as training data for the ML model to learn from. I have enough data to be able to create an ML model that can predict MBRs in the future, and determine when PMT will or will not meet their MBR goal for the day, week, or month.

* **Are you predicting some continuous number, or predicting a binary value?**

I am trying to predict both, this is an example of time-series related problem and the model needs to predict a continues and binary values. I would like the model to predict MBR by either daily, weekly, or monthly. PMT, specifically HUB1 is given a goal each month, so I would like to predict which days, weeks, or months this is likely to happen.

* **What kind of impact do you think it could have?**

There haven’t been any previous attempts to model or predict MBR, so anything that could be done to impact will be a positive in the right direction. In this case would be to reduce PMT’s operational cost.

The opportunity to reduce operational cost $69.00 per bag. Here is a quick breakdown on the effect the model could possibly have:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 245,000.88 | Total number of missed bags in 2016 | |
|  | $69.00 | Liability cost per bag | |
|  | **$ 16,905,060.65** | Total Approximate Annual Operational Cost | |
| 1% Reduction Cost Per year | 3% Reduction Cost Per year | 5% Reduction Cost Per year | 10% Reduction Cost Per year |
| **$169,050.61** | **$507,151.82** | **$845,253.03** | **$1,690,506.07** |

* **What do you think will have the most impact in predicting the value you are interested in solving for?**

At this initial stage and from the customer’s perspective, my thoughts are, number of passengers, number of bags, holidays and seasonality will have impact on mishandled bags, the Mishandled Bag Ratio (MBR).

**Datasets**

* **Description of data set available, at the field level (see table)**

The original data set came in multiple excel files with many tabs, PMT’s reporting is either daily or monthly. I was able to obtain five excel documents, but I merged them together in one big data frame in excel, then converted to CSV format.

* **If from an API, include a sample return (this is usually included in API documentation!) (if doing this in markdown, use the variation code tag)**

**N/A**

**Domain knowledge**

* **What experience do you already have around this area?**

I don’t have any work-related experience in the airline industry, but I have flown in an airplane many times inside and outside the US.

* **Does it relate or help inform the project in any way?**

Yes, we are all customers of the airline industry at some point or another we have flown in an airplane. Based on our flight experience, specifically bag handling makes up for good or bad customer experience. This project given an insight as to how mishandled bags are reported and how they are evaluated.

* **What other research efforts exist?**
  + **Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.**

On my Google search I did not find an actual model or approaches to model “mishandled bags” in an airline. What I found are reports and statistics comparing the MBR across all airlines. When I looked at GitHub there were no MBR models either, most of the ones I ran into had to do with pricing and flight delays.

* + **This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."**

Yes, the team has tried analyzing the MBR using excel. It was a yearly project called “Bootcamp”. They took a more detailed breakdown report of MBR, they looked at buckets for MBR, then assigned to a specific department to make corrections and improvements.

* + **Include a benchmark, how other models have performed, even if you are unsure what the metric means.**

The Bootcamp project had 20% or more success rate. But the project focused more on operational actionable items by department.

I don’t think I can compare my model to this one and make a proper comparison, because the data provided is different than this previous project. Bootcamp was immediate actionable items.

**Project Concerns**

* **What questions do you have about your project?**

What other tools and packages inside python are available to me that can use and explore, that we haven’t learned in class yet.

How many graphs and how much code do we need to use for the project to be complete? What does successful mean from the instructor’s point of view?

* **What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).**

I am not sure if I have enough data? I am requesting to see if I can get 2015 MBR reports, I can only try. Since, I don’t work for the airline industry I am not sure what are other major factors and contributors that affect MBR that are not inside this data set, confounding variables/factors.

* **What are the assumptions and caveats to the problem?**
  + **What data do you not have access to but wish you had?**

The number of total flights handled for the day at HUB1. US weather conditions by day for whole US. Number of employees or crews that work in each department that makes up for bag handling team, such as check-in, transporters, counter, and carousel, TSA rate inspections at this HUB or per flight.

* + **What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample (say, predicting who would win an election)**

This data and model is only restricted to HUB1, and it won’t be predictive or reflective of the overall PMT’s airline baggage performance.

* **What are the risks to the project?**
  + **What's the cost of your model being wrong? (What's the benefit of your model being right?)**

The opportunity to reduce operational cost, or the liability cost to customer is about $69.00 per bag. Here is a quick breakdown on the effect the model could possibly have:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 245,000.88 | Total number of missed bags in 2016 | |
|  | $69.00 | Liability cost per bag | |
|  | **$ 16,905,060.65** | Total Approximate Annual Operational Cost | |
| 1% Reduction Cost Per year | 3% Reduction Cost Per year | 5% Reduction Cost Per year | 10% Reduction Cost Per year |
| **$169,050.61** | **$507,151.82** | **$845,253.03** | **$1,690,506.07** |

* + **Is any of the data incorrect? Could it be incorrect?**

I am confident the data is complete and accurate, PMT’s reporting tools are of high quality. I don’t show missing values or gaps on the data sets provided.

**Outcomes**

* **What do you expect the output to look like?**

Predict MBR behavior, time-series by either month, day, or week. The model will predict day(s), weeks or months MBR goal will be met. At first look holidays, seasonality will come in play in airline industry. I am expecting high bags to passenger ratio to correlate with high MBR (at 1st look anyway).

* **What does your target audience expect the output to look like?**

Here is what I gather from a conversation:

* High passenger counts on low bag to passenger ratio days will produce low MBR - good.
* Tue Wed Thu most likely to make MBR goal vs Fri Sat Sun least likely

Short holidays. Like July 4th - MBR will be low because people fly, and bags don’t get checked-in.

* A0 = Arrival on time

When A0 is 70-100, the MBR will be optimal.

When A0 is 60-70, MBR will start to deteriorate, especially on a big bag day like Saturday.

If A0 is 40-60– We won’t make MBR target on any day.

* **What gain do you expect from your most important feature on its own?**

1 to 3 % improvement per month

 From a conversation – On time plane Arrival to HUB1 was determined to be most impactful.

* **How complicated does your model have to be?**

No boundaries, I will explore all available

* **How successful does your project have to be to be considered a** **"success"?**

The goal for HUB1 is to reduce MBRs year-over-year by 10% or more.

* **What will you do if the project is a bust (this happens! but it shouldn't here)?**

Have the model test in other HUBs, if possible and see how it works. There is nothing official, I don’t work for them, but it will be a good way for me to learn how a DS model project works in real life. If that doesn’t work still, then request additional data, analyze, and readjust model.