FreshBooks Data Scientist - Case Study

January 12, 2016

This case study is designed to assess your analytical methodology, knowledge of modelling techniques and ability to communicate your reasoning. The data attached is not real, however, it is structured in a similar manner to data you may encounter in a SaaS business. We are looking for clear answers that demonstrate a good grasp of common Statistics and Machine Learning scenarios and their real world applications.

The case study is divided into two sections: conceptual and practical. For the conceptual section we are looking for short, qualitative explanations that show not only ability to tackle a variety of problems, but also to justify and explain your approach. For the practical section we are interested in gauging your ability to solve real data problems. You will **not** be judged on the performance of your predictions, however a reasonable level of accuracy is expected.

Please explain your reasoning whenever appropriate and note any assumptions you make. **You are not required to complete the case study,** answer as many questions as you can (or want).

1 Conceptual

1.1 Probability and Statistics

In this section we tackle from a high level the two main approaches to statistical inference: Frequentist and Bayesian.

1.1.1 The Likelihood Function

1. Perhaps the most common method of fitting models to data is by maximizing the likelihood function. In this question we will discuss broadly why maximizing the likelihood makes sense, its advantages and disadvantages, and how to avoid common pitfalls.

    1. Explain why estimating parameters of a model via the maximum likelihood procedure makes sense.
    2. What are some advantages and disadvantages of estimating parameters by maximizing the likelihood?
    3. Given model parameters θ and data $\mathcal{D}$, what are some common methods to avoid overfitting? Explain briefly how they work.

1.1.2 Bayesian Analysis

2. In this question we will address two other ingredients (in addition to the likelihood) present in Bayesian statistics: the prior and the posterior distributions. A challenge question (closely related to Bayesian analysis) is also included.

    1. What is the posterior distribution?
    2. Explain what a prior is and what impacts it can have on the posterior distribution.
    3. You are assigned the task of calculating the area under an unknown curve $f(x)$ in the interval $x \in (-10, 10)$. You are not able to evaluate the curve at an arbitrary point x, instead, given a tuple (x_0, y_0) you are only able to tell whether $y_0 > f(x_0)$ or not. At your disposal you only have a random number generator that gives samples uniformly between zero and one ($s \sim U(0, 1)$).

Additionally you are told that the curve is guaranteed to always be greater than zero and less than two ($0 < f(x) < 2$, $\forall x$). Explain how you would approach this problem.

1.2 Machine Learning

In this section you are asked to outline possible ways of approaching common scenarios where machine learning can be applied in a business setting. We are looking for short, qualitative answers that showcase not only your general knowledge, but also your ability to convey information clearly and effectively.

3. You are given a dataset with 100k **labeled** samples and 200 features, all labels are binary, how do you proceed if:
    1. For some of the features a large proportion of the data is either missing or corrupted.
    2. Your samples follow a natural order and strong time dependence is expected.
    3. Many features are categorical, with some having a large number of different categories.
    4. What would change (if anything) if the dataset had, instead, 2000 samples and 1000 features?
4. The Sales Department is interested in identifying possible leads for a new upsell campaign happening early in the customer's life time. The number of new customers joining the company each day is too large for the sales team to contact directly, therefore it is important that their time is focused on higher quality leads (more likely to take the upsell). You are tasked with building a model that feeds the sales team with the leads they should work on and the order that they should contact them. Available is all historical data of older customers and whether they would be considered good or bad leads.
    1. Discuss on a high level how you would approach this problem.
    2. What metric would you use to access the performance of your model? Why?

2 Practical

In this section you are asked to build models, make predictions as well as recommendations to stakeholders. Together with this case study you should have received three CSV files: dataset_1.csv, dataset_2_train.csv and dataset_2_validation.csv (an explanation for the type of data contained in each file in present at the end of the case study).

5. A marketing channel manager wants to know if he should distribute his budget according to the day of the week (allocating more money to days that convert better), help him answer this question.
    1. Refer to dataset_1.csv where you can find information on the number of leads and subsequent converted leads by date.
    2. Explain your assumptions and how you decided to model the problem.
    3. Make a recommendation to the channel manager.
6. Understanding and predicting churn is extremely important for any subscriber based business. It can help with Marketing, Sales, Customer Support and even Product Development strategy. In this question you are asked to build a model to predict which customers will churn in the next month as well as the probability or churning in each of the next 12 months. For this question, please refer to the dataset_2_train.csv and dataset_2_validation.csv files; explanation for each file's structure is available below.
    1. Use dataset_2_train.csv to build a model that predicts which customers will churn. Include your model as well as a short explanation with the answers to this case study.
    2. Make predictions for all users present in dataset_2_validation.csv.
    3. Use dataset_2_train.csv to build a model that predicts the likelihood of churning at every subsequent month for the next 12 months. Include your model as well as a short explanation with the answers to this case study.
    4. Make predictions for all users present in dataset_2_validation.csv for the next 12 months. 5 Include your predictions with the answers to this case study.

3 Dataset Descriptions

1. **dataset_1.csv:**
    1. start_date: date
    2. day_of_week: corresponding day of week
    3. leads: Number of leads that signed-up at this date
    4. converted_leads: Number of leads (from the ones that signed-up) that moved to a paid solution
2. **dataset_2_train.csv:**
    1. id: Unique identifier
    2. f1-f16: Numerical features
    3. f17-f18: Categorical features
    4. churned_within_30_days: Whether or not the customer cancelled his (hers) subscription within the next 30 days
3. **dataset_2_validation.csv:**
    1. id: Unique identifier
    2. f1-f16: Numerical features
    3. f17-f18: Categorical features