Personal Financial Data Mining Update

Version 1.1

Mark Kurzeja

Abstract

For the last five-plus years, as a part of my monthly budgeting, I've been collecting information about every single dollar that I have ever spent during my times in junior year of college, senior year of college, my time living in Manhattan, and through my first and second year of graduate school. During this time, I found that budgeting and the prediction of even my own personal finances have been incredibly difficult and sometimes downright frustrating. While most of the time, the consequences for not budgeting exactly can be relatively minimal, there have been moments in my life where prediction errors have resulted in the times that have caused financial difficulties. Because of the emotional and financial difficulties that these shortfalls can cause, I have tried over the years to create simple metrics to predict upswings in spending and prevent myself from overspending. My goals for this project are twofold, 1) I would like to create better predictions of my own financial spending on variable items such as eating out, entertainment, and other expenses that are within my control, and 2), I would like to synthesize these into in index or prediction scale that I can use to monitor my current financial state which acts as a forward indicator of potential financial shortfalls in the future.

1 Introduction

1.1 The Problem

Managing one's own finances can be hard. It is no secret that the average American is rather terrible with money. It is estimated that as much as 20% of the population has a negative net worth (1), and during my time working in the financial services industry, I've seen the results of many poor financial decisions completely wreck people with incomes tens of times what the average American makes. I find myself to be fortunate that I have been able to maintain somewhat strict rules for myself, and for the past five or so years I have been tracking every single dollar that I've ever spent by hand. I've done this via the use of a budgeting app called YNAB. There have been times that I have been very good at estimating my future expenses, and when this happens I'm better able to predict my financial picture during the coming months. Especially whenever I have been in college, and in grad school, there have been times that I've had to go almost a full year without a paycheck. Whenever I lived in Manhattan, large expenses such as rent meant that if I miscalculated my other expenditures, I would not have enough money to do simple things such as make it back home for the holidays or see my girlfriend at the time.

1.2 First and Second Order Goals

Sometimes however the consequences of not predicting your financial picture accurately can be devastating. we live day today not really knowing what tomorrow will hold, but whenever it comes to money, ignorance is anything but bliss. Errors in estimation can result in severe financial difficulties for some, and my hope with this project is to be able to provide a far more accurate estimate of my finances in the future using information about my spending patterns today and my spending patterns in the past. My hope is that my research into my own finances will be generalizable into that of my family which actually tracks all their finances in the exact same manner that I do. I hope that I can expand the system and one day I actually hope to own my own Financial advisory firm. My research into this topic, and the methods that I devised, I hope will one day be useful in predicting financial inflows and outflows in an effort to help people get a better track of their financial futures and ensure that they are able to achieve their goals despite the fact that the world is inherently chaotic, messy, and sometimes downright unpredictable.

2 Problem Definition and Data

My goal for this project is twofold:

1. I want to create a system of models that are able to better predict my reoccurring variable spending habits using his little of data as possible to get the most accurate predictions on a monthly basis. I want to use as little of data as possible in order to ensure that the system can work for other people who do not have a large financial history such as I do. And I want the system to be is accurate as it possibly can given the information that I put in.

2. I want to distill this prediction mechanism into an index they can track how people are doing overtime. Like a performance gauge, I want someone to know when they are doing well and are spending under what we would expect them to be spending at any given moment, and I want them to know whenever they are spending poorly and they need to curtail their current expenditures in order to get back on track. I think the simple metrics such as financial ratios aim to do this four people in a hand calculation sort of way. My aim is to make something that someone can track with their eyes so that they can see if they are doing well and so that they can manage their finances eventually on their own without the need of a financial adviser to tell them if they are doing well or not.

This project is not solely focused on predictions for one month out. This project is also not simply about forecasting. The models that I wish to build I want to be able to derive insights from. I want to be able to take the results and put them in the words and explain to people why the model may have over or under predicted in a given month, and why it thinks that its predictions are going to be on track.

My evaluation metrics will include asymmetric L1 loss (the losses resulting from over-prediction and a shortfall of money are far less severe than the losses resulting from foregone investment interest resulting from under budgeting and capital surplus). I want to run a conjoint analysis to determine what loss I can create that best aligns with how I experience loss. For this, I will place a large number of scenarios in front of myself and I want to evaluate which I would prefer in a given circumstance. For example, would I prefer to over budget by $100 or would I prefer to under-budget vacation by $80 and not be able to go out to dinner on the last night? Answers to questions such as these will be very useful in determining how I should assign loss measures and to which categories of my budget I should be the most focused on.

2.1 The Data Set

I have every single dollar tallied that I have spent in the past few years in my budgeting program. The data for a transaction includes the following:

Date The Date the transaction took place

Account From which of my personal accounts did the money come from? Checking, Savings, Etc?

Master Category A high-level category assigned by myself. This includes things like is this a Housing Expense? Is this a Leisure Expense? Is this a Financial Expense? etc.

Sub Category This is a more granular category that is meant to break up the master categories. If something is a housing expense, is it for rent, or household items, or for a security deposit?

Payee Who was the transaction to?

Outflow If the item was an outflow, what was the amount?

Inflow If the item was an inflow, what was the amount?

Memo User-generated description of the expense

Further to each of the data points above, I have the total amount spent in each category during the last five years and I can see how I planned for expenses in the future using the data from my budgeting values.

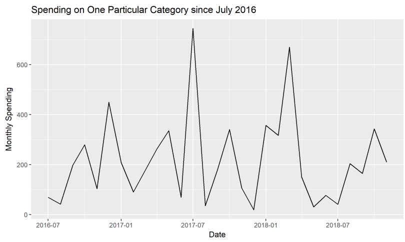

Figure 1: Monthly aggregation of spending for a particular category

Some of the challenges that I see from working with this data include:

Sparsity The fact that I only record a transaction once it happens and not when I expect it to occur means that for certain categories that are less frequent, there could be very little data predicting incidence in the future. For instance, I have only taken two vacations in the last three years which means that months were vacations are in the mix will be mostly zero expenditure with some very large outliers

Zero-inflated models To the sparsity point, most of the entries in the data are zero. Since some events only occur every other week or similar, I have to model both the non-occurrence of an event as well as the expectation when it does occur.

Privacy This model will have to be anoymonized to a certain extent since this is a complete account of my financial picture. As a result, some censoring or name replacement may be necessary depending on the features that I find to be useful in predicting outflows.

The dataset that I am working with is 2625 transactions spread across several years worth of data. I am currently working on integrating older data as well which will get the dataset to be around 4000 transactions in total, but as of right now, the dataset will remain at 2625 transactions. Since my problem inherently is a time-series prediction problem using sparse data, it is a continuous regression problem and with transaction data as granular as daily, I have a lot of features that I have been working to engineer for the dataset including running averages, lag period prediction, and other methods that may improve the prediction on the coming months.

Some of the most frequent features of the data include severe outliers and large periods of days where there are no transactions to note. When I lived in New York, there were entire month periods where I didn't eat much at all at a restaurant, and as a result, those months appear as dead-months in the dataset. Similarly, if I was out for my girlfriends birthday at the time, then I would see that this shows up as a clear outliers in the dataset. It is these sporattic measurements that make this a difficult prediction problem even for mean estimation and this is something that I am working on trying to figure out.

The data is as clean as my recording has been over the last three years. Since this register is tied to the cent to my bank accounts every week, I know that the data is good and it should be very clean, but the only information that I have on each transaction, other than the parties involved and the month, are the comments that I leave and there are many of them without a comment in the dataset. An example of the data is contained in Table 1.

3 Related Work

There has been a lot of work on personal forecasting but most of it, to maintain simplicity, has been based off of using simple averages and running averages to predict the spending in the coming months.

More advanced methods of prediction have, to my knowledge, never been used on personal financial data in this fashion before, and so I am to bring the more advanced analytical toolbox to the personal financial space in an effort to improve the forecasting accuracy and budgeting prediction. While methodologies are probably implemented in the private wealth space for ultra-high net worth clients, the people that need this prediction setup the most are the people who don't have the money to afford the inaccuracies that budgeting can allow in the first place.

The prediction problem for time series is well studied however, and from that perspective there are a variety of methods that have been brought to the table. Anything from ARIMA modeling to Facebook's Prophet have been found to be useful. Gaussian Processes are something that may be useful to get variance estimates for each of the predic-

Account	Date	Payee	Master Category	Sub Category	Memo	Outflow	Inflow
Capital One Quicksilver	8/13/2016	Diner Airport	Everday Wants	Partying, Drinks, and Food	Food airport	$24.04	$0.00
Capital One Quicksilver	8/13/2016	Quik Fill	Everday Wants	Partying, Drinks, and Food	Trip tea	$1.38	$0.00
Cash In Wallet	8/20/2016	Jakes Saloon	Everday Wants	Partying, Drinks, and Food	Justin and Pete	$27.00	$0.00

Table 1: Three Data points from the data set

tions, and other models that allow for smoothing of noisy data may be useful as well. ASAP, which we saw in class, is something that may be able to take out a large variety of the variation that financial data often brings about as well.

In (2), the authors used SVM, training on several properties of the time series up until this point, to predict how the stock will develop in the coming months. For this projects, I could use similar features such as volatility in the previous periods as well as skew and other technical oscillators borrowed from technical finance such as the Relative Strength Index and the Commodity Channel Index to use as training features and then run supervised machine learning algorithm on the data. In this way, I could hope to get away from the nature of the time-series modeling I have been doing in order to begin creeping into the more classical approaches to prediction treating the new monthly average as a predictive of the previous months.

In (3), the authors used extreme-value modeling to come up with predictions of tail-risk in their financial modeling. Looking at my project from this lens, this is the measure that I am trying my hardest to estimate. Seeing as most people don't really care about the expected value of their spending, but rather would like to know the maximum for risk modeling, this paper actually may have changed my problem statement for some of the modeling a little bit. I would like to explore how to use these models to better estimate tail risk, and I want to build a model that is able to account for the tail risk of my spending.

4 Methodology

Some of the methods that I plan to try for the forecasting piece of the assignment are as follows:

1. As a baseline, how good / bad are running averages at predicting the future for financial data?

2. ARIMA as a baseline

3. Baysian Gaussian Processes as a means of modeling time series data with uncertainty estimates

4. Baysian Hierarchal Modeling to predict the distributions of 1) the occurrence of a day with a particular type of spending. i.e. did I buy groceries today? and 2) if spending occurs, what is the distribution of the spending? i.e. I'm going to the grocery store, what am I going to spend?

5. Facebook's Prophet - I've heard great things about the success of this system in predicting time series. I'm curious how well it will perform on the sparse data that I have

6. Smoothing techniques like filtering - sometimes noise reduction allows one to see the forest through the trees - I'm hopeful that something like ASAP or DFT will allow the noise that is present in a time series to be reduced for financial data enough to make robust predictions.

7. Sequence processing such as a hidden Markov model may be very helpful as well and it is something that I would love to try

Some of the methods that I plan to try for the metric piece of the assignment are as follows:

1. Monthly Spending Index - at any given time, how much do I expect to spend in the next 30 days? This metric could be updated in real time to show an index of when I have been overspending or underspending

2. Bullet charts to show good-caution-danger zone estimates for certain categories. Popularized by Stephen Few, these are great little charts and are highly dense information wise and can be very quickly made into a dashboard `https://en.wikipedia.org/wiki/Bullet_graph`

3. Pain Index - Given the conjoint information from the beginning assessment, how likely are you to experience pain in the form of under-budgeting within the next month? Tracking this over time will let someone know how close they typically are to their goal and provide course correction in the event that they are unaware that they are not on track.

I plan to follow the methods plan that I have suggested above to identify which of the methods it the best at predicting financial data in the future. I would like to see how different methods compare and if I can gather anything out of the various methods by gaining predictive power or reducing the variance of the estimates.

5 Evaluation and Results

From above: My evaluation metrics will include asymmetric L1 loss (the losses resulting from over-prediction and a shortfall of money are far less severe than the losses resulting from foregone investment interest resulting from under budgeting and capital surplus). I want to run a conjoint analysis to determine what loss I can create that best aligns with how I experience loss. For this, I will place a large number of scenarios in front of myself and I want to evaluate which I would prefer in a given circumstance. For example, would I prefer to over budget by \$100 or would I prefer to under-budget vacation by \$80 and not be able to go out to dinner on the last night? Answers to questions such as these will be very useful in determining how I should assign loss measures and to which categories of my budget I should be the most focused on.

In math, one such metric would be as follows (let R be the realized amount of spending and B be the amount Budgeted):

$$Loss(R, B) = w_1 \max(R - B, 0) + w_2 \max(B - R, 0) \tag{1}$$

Where, if $realized > budgeted$, then we have overspent and w_1 will drive our loss and the converse holds true for w_2. Given our discussion from before, it makes sense that $w_1 \geq w_2$ since the pain of having a shortfall, for most, will be greater than the gain that one experiences from optimizing their free cash flow. However, when I run the conjoint analysis, I hope to find out if this is really the case.

The baseline metrics that I will use are both very common in the space:

1. Running (or Simple) Average - Simple means and measures of center are the most common implementation for many of the prediction metrics from month to month. This is the method that is used in my budgeting program currently, and its inadequacy is the main reason why I thought of this project.

2. ARIMA - this is probably the most sophisticated model that most anyone who works with basic financial data uses since it is widely implemented and this is usually the most advanced (and only) mention of time series in most business schools[1].

Looking at the Baseline models, there is still lots of improvement to do to determine how the models will perform, but with an L_1 loss all above 150, I think that the unweighted mean and the ARIMA models are going to be good baselines to begin with. I have carried out the following baseline models with their results:

- Basic lagging mean: This method used the running equal-weight mean with lag periods 1:15. If $lookback = 0$, this is equivalent to using the previous month's value to predict this months value, and a lag of fifteen would use the average from the previous fifteen months to predict this month. The error is printed out in terms of the average L_1 error that each method produced. See figure 2 for the visualization of the errors We can see that the highest lags had the lowest errors

- ARIMA Modeling: The ARIMA error is expressed in Figure 3. We can see that the simplest model (one without a linear trend or seasonality) has the lowest error on the chosen problem

[1] During my time in Ross, this was only mentioned in an advanced analytics course at the 400 level and was a highly non-standard part of the curriculum. During my time in banking, most forecasting was done at a far more micro level using correlations and so large scale time series analysis were left to the research economists who used mean expectation algorithms for stability.

	Type	error
1	Mean(lookback = 0)	225.68
2	Mean(lookback = 1)	203.28
3	Mean(lookback = 2)	183.10
4	Mean(lookback = 3)	191.36
5	Mean(lookback = 4)	184.09
6	Mean(lookback = 5)	170.71
7	Mean(lookback = 6)	168.43
8	Mean(lookback = 7)	160.13
9	Mean(lookback = 8)	157.74
10	Mean(lookback = 9)	160.21
11	Mean(lookback = 10)	160.86
12	Mean(lookback = 11)	159.83
13	Mean(lookback = 12)	159.31
14	Mean(lookback = 13)	159.76
15	Mean(lookback = 14)	157.38
16	Mean(lookback = 15)	158.63

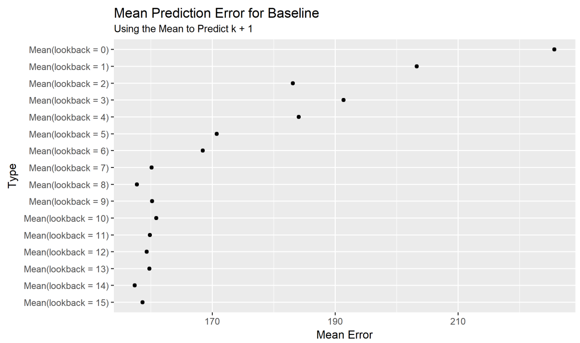

Figure 2: Average L_1 error for the lagged mean models

Expansions on this, to be implemented shortly, include different time period aggregations (maybe weekly) or to use other methods of predicting the outflows in a month (Monte Carlo or other ensemble methods to build up months one day at a time.)

6 Discussion

With an average error for the simple mean calculation of over 150 and over 160 for the best ARIMA model, the baseline performs rather poorly but this is something that I think that could be useful for measuring the success of future methods. In short, if you ran this model, and asked it what you would spend in the next month, it would give you an estimate with an average L_1 error of \$150 which tells us that if it predicts \$400 in spending on food, that the error bars are from \$100 to \$700 which make the estimates next to useless.

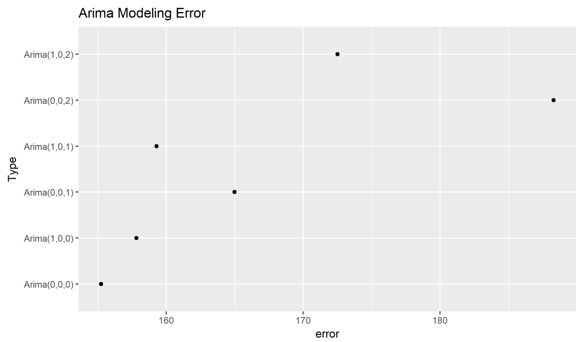

Figure 3: Anova Error for R Present Models

This is a good baseline to begin off of, because it tells us that the mean function of the past data is not the most important thing in predicting this data on a pure-monthly basis, and this tells us that methods will inherently need to work with more granular aggregations (which are possible since I have the daily data) in order to gain useful insights into the modeling of the expected value.

7 Work Plan

My plan for this project roughly follows the following guideline:

1. Gather the data from my budgeting program. I already have a workflow for this This has been completed

2. Determine which categories are worth modeling - I'm looking at any variable expenses as being the most useful thing to model. Fixed expenses like rent are contractual and can be modeled statically I have decided to spend most of my time modeling categories like "Spending money" which is a catch-all for random things like house supplies and school materials and "Food and Drinks" which includes every single meal that I have eaten out that does not include the meals that I make with my own groceries as well as "Groceries" which is my tracking of my grocery expenditures.

3. Begin building look-ahead prediction models and comparing them.

   (a) Start with base models. Assess prediction accuracy This has been implemented as of this writing using the baseline for the means as well as the ARIMA model

   (b) Work up to more advanced models
   (c) Try to combine results using stacking to improve prediction error I don't think that this step will happen due to the modeling complexity already
   (d) Reduce data-size as much as possible to see how much data we need for "good enough" accuracy This is an implicit part of the error analysis and will probably be included as a result

4. Begin looking at metrics that assess the health of a persons spending as a means of visualizing progress. This is the part of the project that is the most TBD due to the fact that the timing is going to be quite intensive to model out all of the other parts and get the performance of the models where I would like to. This, I think, is the most useful part of the project though for my purposes and it is something that I would like to focus on if I get the time.

   (a) Track financial ratios and see if they are "useful"
   (b) Look at smoothing procedures to see if they add information
   (c) Look at simulation based visualizations to see if they add anything interesting

Acknowledgments

My brother, an accountant, and I as well as several friends have discussed beginning a financial advisory firm focusing on tax arbitrage and investing advice based on tailored risk analysis. My background in finance and financial modeling heavily informs my views of financial predictions in general, and I know how noisy the real world is with this kind of data. Having always been interested in personal finance, this aspect of modeling out my finances is something I have attempted in the past using various simulation based methods, but I have never taken a deep dive into quantifying error nor coming up with a system for prediction and analysis. I hope to expand this for the ambitions of a past self and for the benefit of a potential business

References

[1] BNP Paribas - During my time at BNP Paribas in New York, this was often a statistic that was used in client meetings and for presentations.

[2] Kim, Kyoung-jae. Financial Time Series Forecasting Using Support Vector Machines. NeuroImage, Academic Press, 13 May 2003, www.sciencedirect.com/science/article/pii/S0925231203003722#SEC

[3] McNeil, Alexander. Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach. NeuroImage, Academic Press, 1 Dec. 2000, www.sciencedirect.com/science/article/pii/S0927539800000128.