A Project/Dissertation Review-1 Report

on

New Customer Evaluation Metrics: Clumpiness

Submitted in partial fulfillment of the requirement for the award of the degree of

Bachelors Of Engineering
Computer Science

Under The Supervision of Prashant Johri : Professor

Submitted By Udbhav Singh Patwal
19SCSE1180060

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
11, 2021

Abstract

Previously we have encountered various problems in evaluating the customer metrics and taking out the customer lifetime value which is very important to the firms all around the world. The RFM metrics currently used in the market has been over 40 years old and there is a variety of changes happened to calculate the probabilities of a desired customer. Thus our approach adds up a new metrics in line to fees with the new era. However, since many concerns have been raised about the low power of existing “hot hand” significance tests, we propose a new class of clumpiness measures which are shown to have higher statistical power in extensive simulations under a wide variety of statistical models for repeated outcomes. Finally, an empirical study is provided by using a unique dataset obtained from Hulu.com, an increasingly popular video streaming provider. Our results provide evidence that the “clumpiness phenomena” is widely prevalent in digital content consumption, which supports the lore of “bingeability” of online content believed to exist today

To evaluate the metrics we have used the help of the Machine learning Models. Python has been used as a base environment in Google Colab. Results are based on series of evaluation of customer’s value in the next half year time frame.

Data clumpiness is usually considered as irregular cluster(s) of activity gathered together, and many models have been developed in the literature to analyze those “clumpy” data, such as the hidden Markov, self-exciting point process and autoregressive conditional duration models. However, there is a lack of nonparametric, exploratory data analysis measures of clumpiness. In this paper, we provide researchers with a class of new measures to assess clumpiness. This class of measures is useful both as an exploratory data analysis tool and as a way to assess whether a model has fully captured the clumpiness in the data, e.g., by using the measure to carry out a posterior predictive check (Gelman et al., 1996).

List of Tables

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Table of Contents

CHAPTER-1
Introduction

One of the most established practices in the field of marketing and customer valuation is to summarize a customer using what's called RFM segmentation — regency, frequency, monetary value — which means I take everything I know about my customer and I compute just three simple numbers: How recently did they buy? How frequently do they buy? And when they buy, how much money do they spend? … It's the basis on which most companies decide who are the valuable customers and who are the non-valuable customers. RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns. The system assigns each customer numerical scores based on these factors to provide an objective analysis. RFM analysis is based on the marketing adage that "80% of your business comes from 20% of your customers."

RFM analysis ranks each customer on the following factors:

Recency. How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more likely to purchase or use the product again. Businesses often measure recency in days. But, depending on the product, they may measure it in years, weeks or even hours.

Frequency. How often did this customer make a purchase in a given period? Customers who purchased once are often are more likely to purchase again. Additionally, first time customers may be good targets for follow-up advertising to convert them into more frequent customers.

Monetary. How much money did the customer spend in a given period? Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business.

RFM analysis scores customers on each of the three main factors. Generally, a score from 1 to 5 is given, with 5 being the highest.

CHAPTER-1
Introduction

Various implementations of an RFM analysis system may use slightly different values or scaling, however.

The collection of three values for each customer is called an RFM cell. In a simple system, organizations average these values together, then sort customers from highest to lowest to find the most valuable customers. Some businesses, instead of simply averaging the three values, weigh the values differently.

For example, a car dealership may recognize that an average customer is highly unlikely to buy several new cars in a timeframe of just a few years. But a customer who does buy several cars -- a high-frequency customer -- should be highly sought after. So, the dealership may choose to weigh the value of the frequency score accordingly.

RFM analysis is also valuable for organizations that do not sell products directly to customers. Nonprofits and charities can use RFM analysis to find the best donors, for example, as those who have donated in the past are more likely to donate again in the future.

Lastly, businesses that do not rely on direct payments from customers may use different factors in their analysis. For example, websites and apps that value readership, number of views or interaction may use an engagement value instead of monetary value to perform an RFE (recency, frequency, engagement) analysis instead of a standard RFM analysis using the same techniques as the latter.

One of the most settled practices in the field of promoting and client valuation is to sum up a client utilizing what's called RFM division — rule, recurrence, money related worth — which implies I take all that I am familiar with my client and I register only three basic numbers: How as of late did they purchase? How often do they purchase? Also when they purchase, how much cash do they spend? … It's the premise on which most organizations conclude who are the significant clients and who are the non-important clients.

CHAPTER-1
Introduction

RFM examination is an advertising method used to quantitatively rank and gathering clients dependent on the recency, recurrence and financial complete of their new exchanges to recognize the best clients and perform designated showcasing efforts. The framework doles out every client mathematical scores dependent on these variables to give a goal investigation. RFM examination depends on the advertising maxim that "80% of your business comes from 20% of your clients."

RFM investigation positions every client on the accompanying elements:

Recency. How ongoing was the client's last buy? Clients who as of late made a buy will in any case have the item at the forefront of their thoughts and are bound to buy or utilize the item once more. Organizations regularly measure recency in days. In any case, contingent upon the item, they might quantify it in years, weeks or even hours.

Recurrence. How regularly did this client make a buy in a given period? Clients who bought once are regularly are bound to buy once more. Furthermore, first time clients might be great focuses for follow-up publicizing to change over them into more successive clients.

Financial. How much cash did the client spend in a given period? Clients who burn through large chunk of change are bound to burn through cash later on and have a high worth to a business.

RFM investigation scores clients on every one of the three principle factors. By and large, a score from 1 to 5 is given, with 5 being the most elevated. Different executions of a RFM investigation framework might utilize somewhat various qualities or scaling, nonetheless.

The assortment of three qualities for every client is called a RFM cell. In a straightforward framework, associations normal these qualities together, then, at that point, sort clients from most elevated to least to track down the most important clients A few organizations, rather than basically averaging the three qualities, gauge the qualities in an unexpected way.

CHAPTER-1
Introduction

For instance, a vehicle sales center might perceive that a normal client is profoundly far-fetched to purchase a few new vehicles in a time period of only a couple of years. Be that as it may, a client who purchases a few vehicles - - a high-recurrence client - - ought to be profoundly pursued. Thus, the showroom might decide to gauge the worth of the recurrence score appropriately.

RFM investigation is additionally important for associations that don't sell items straightforwardly to clients. Philanthropies and good cause can utilize RFM examination to track down the best benefactors for instance, as the people who have given in the past are bound to give again later on.

Finally, organizations that don't depend on direct installments from clients might involve various variables in their examination. For instance, sites and applications that esteem readership, number of perspectives or connection might utilize a commitment esteem rather than money related worth to play out a RFE (recency, recurrence, commitment) investigation rather than a standard RFM examination involving similar strategies as the last option.

One of the most settled practices in the field of advancing and customer valuation is to summarize a customer using what's called RFM division — rule, repeat, cash related worth — which infers I take all that I know about my customer and I register just three essential numbers: How actually did they buy? How frequently do they buy? Additionally when they buy, how much money do they spend? … It's the reason on which most associations close who are the critical customers and who are the non-significant customers.

RFM assessment is a promoting strategy used to quantitatively rank and assembling customers reliant upon the recency, repeat and monetary complete of their new trades to perceive the best customers and perform assigned displaying endeavors. The structure gives out each

customer numerical scores reliant upon these factors to give an objective examination. RFM assessment relies upon the publicizing saying that "80% of your business comes from 20% of your customers."

RFM examination positions each customer on the going with components:

Recency. How progressing was the customer's last purchase? Customers who actually made a purchase will regardless have the thing at the cutting edge of their considerations and will undoubtedly purchase or use the thing again. Associations routinely measure recency in days. Regardless, dependent upon the thing, they may measure it in years, weeks or even hours.

Repeat. How consistently did this customer make an up front investment a given period? Customers who purchased once are consistently will undoubtedly purchase again. Besides, first time customers may be extraordinary concentrations for follow-up publicizing to change over them into more progressive customers.

Monetary. How much money did the customer spend in a given period? Customers who consume huge load of cash will undoubtedly consume cash later on and have a high worth to a business.

RFM examination scores customers on all of the three guideline factors. All things considered, a score from 1 to 5 is given, with 5 being the most raised. Various executions of a RFM examination structure may use fairly different characteristics or scaling, regardless.

The variety of three characteristics for each customer is known as a RFM cell. In a clear system, affiliations ordinary these characteristics together, then, sort customers from generally raised to least to find the main customers A couple of associations, rather than essentially averaging the three characteristics, check the characteristics suddenly.

For example, a vehicle deals focus may see that a typical customer is significantly implausible to buy a couple of new vehicles in a time-

frame of two or three years. In any case, a customer who buys a couple of vehicles - - a high-repeat customer - - should be significantly sought after. In this manner, the display area may choose to check the value of the repeat score suitably.

RFM examination is furthermore significant for affiliations that don't sell things clearly to customers. Philanthropies and great objective can use RFM assessment to find the best advocates for example, as individuals who have given in the past will undoubtedly give again later on.

At last, associations that don't rely upon direct portions from customers may include different factors in their assessment. For example, locales and applications that regard readership, number of points of view or association may use a responsibility regard rather than cash related worth to play out a RFE (recency, repeat, responsibility) examination rather than a standard RFM assessment including comparable procedures as the last choice.

CHAPTER-2
Literature Review

The RFM model is fundamentally built using principles of data-driven marketing. Data-driven marketing has fundamentally transformed how marketing works ever since its inception, as it allows the analysis of large sets of customer data like never before. This has led to increased accuracy in understanding customers and enhanced ability to creatively customize messaging. The rise of automation in marketing technology has led to increased granularity and personalization, leading to enhanced relevance of each brand message.

RFM traces its origin back to 1995 when it was cited by Bult and Wansbeek in an issue of Marketing Science. Used in the context of direct mail, it showcased how the three criteria could be used to better estimate demand, reducing costs on printing and shipping, leading to enhanced returns. With the rising sophistication of computing power, RFM has become easier to apply in businesses due to the computerized customer histories of today.

The RFM model is linked with the famous Pareto Principle, which says that 80% of total results are driven by the top 20% causes. When applied to marketing, it means that ***80% of your total sales are likely to come from your top 20% of customers***. Regular customers will always be high contributors to business revenue, and hence the retention of those customers is highly critical for business performance.

Small businesses constantly face the pressure of acquiring new customers, which define their growth and trajectory, and are prone to spending high amounts of money to acquire them. A business cannot sustain itself without customers, and while acquisition is a critical part of business strategy, retention plays a bigger role in ensuring high returns for the business. Customer retention depends on customer satisfaction with the product, service provided by the business, and the interactions the customer has with the business,

hence making them feel valued.

The digital world is a buyer's market, with a plethora of options available to a user at their fingertips. Brands are constantly jostling and fighting for a share of the customer's wallet and attention. In such an atmosphere, understanding customer behavior and segmenting them into distinct groups, help businesses focus their marketing efforts on relevant customers.

With the power of social media at their fingertips to express displeasure and

the ease of choosing alternatives, customer expectations regarding the quality of brand interactions are high. Hence creating relevant and personalized messaging, tailored to user behavior has become the norm.

Personalization is one of the major benefits of RFM, as it not only allows you to target different customers with varying but equally relevant messaging, but also gives businesses the ability to recognize changing patterns of user behavior through the capture of RFM data, and move the customers to other segments if required.

Through RFM, businesses can recognize and focus on converting critical customer segments like customers on the verge of churning out to becoming active customers, and also encouraging customers who are loyal to the brand to become ardent followers. By minimizing the waste of resources through effective targeting, RFM helps businesses utilize their marketing budgets wisely and effectively, while also increasing the overall impact of marketing on the business.

Media selection

Once RFM analysis is completed, there is an increased understanding of what the user needs most from your brand, and based on behavior, when are they likely to interact with you. A differentiated media strategy, combining multiple formats and mediums, for varying durations, can be created to target different segments based on their characteristics.

Messaging

RFM analysis allows you to create customized and personalized messaging, and this can be used to streamline the various messages you send to a specific customer and continue sending messages of only a particular type, thereby reducing the chance of dissatisfaction or annoyance, and create higher customer satisfaction.

New launches

RFM allows you to recognize your most valuable and least valuable customers, and during the launch of a new product, the Champion

customer can be engaged in a way that creates high WOM, which positively impacts product perception amongst other customers, leading to greater awareness and eventual purchase.
What is clumpiness? To the easygoing watcher, a grouping of (occurrence) information is called clumpy
when burst(s) of exercises or clump(s) of occasions are noticed, however this is anything but a legitimate factual depiction. In the language of measurements, clumpiness shows non-steady affinity, explicitly impermanent heights of inclination—for example periods during which one occasion is more
liable to happen than the normal level. Subsequently, it imparts practically a similar plan to the "hot
hand" impact. Therefore, we will utilize a similar functional meaning of clumpiness as in the
"hot hand" writing—sequential reliance or non-stationarity, however decipher them in an alternate
way.
In the "hot hand" writing, two normal comparing principles to recognize clumpiness
are: (a) that a player ought to act in a manner where "achievement breeds achievement" and (b) there
5
ought to be dashes of achievement in which execution has been raised that stand apart from streaks
because of karma. (a) is tried by inspecting restrictive probabilities and (b) is tried by analyzing
the number and length of runs. Allow us to take a gander at them individually. What's the significance here by "achievement
breeds achievement"? Does it basically mean one result of progress would upgrade inclination of
the following preliminary? What might be said about two results of progress or three results? There is dependably a
challenge in picking the number of past results ought to be incorporated and which examples of
history ought to be looked at.
Concerning second norm, the inquiry we raise is whether a cluster of occasions is
comparable to a run? Is clumpiness just connected with back to back victories? Does "being
hot" have to do with the term and recurrence of streaks? Not really. Clumpiness as it were
proposes that the inclination ought to be bigger than the normal level over some period(s) of
time, yet it doesn't ensure that victories need to happen sequentially. In this way one disappointment
inside a grouping of triumphs doesn't demonstrate that the "hot" period essentially finished on
that day. Two sequential triumphs ought not be dealt with too

uniquely in contrast to the situation when
they are isolated by one disappointment.
Those disadvantages heuristically clarify why existing measures need power in recognizing the
"hot hand" impact. Nonetheless, in the event that one considers the "hot hand" or clumpiness as not being straightforward
successive reliance between preliminaries, or an amazingly improbable streak—then, at that point, by what other means ought to it be depicted and estimated? We examine this next.
3.1 Desired Properties
Having expressed the issue, we presently continue to construct new measures. The typical way is to
propose some action and afterward present its properties; rather we do the converse way — come
up with a rundown of properties which a sensible measure ought to have, and afterward present new
measures which have those properties. The proposed properties are as per the following:
• Minumum : The action ought to be the base assuming the occasions are similarly separated.
6
• Most extreme : The action ought to be the greatest assuming every one of the occasions are assembled.
A
B
• Progression : Shifting occasion times by a tiny sum should just change the
measure just barely.
• Union : As occasions move closer(away), the action ought to increase(decrease).
ti
ti+1 ti+2
max min max
C
The Minimum and Maximum properties are clear and straightforward. Observationally, paying little mind to the quantity of occasions, it is normal to think about the instance of all occasions grouped
together as the most clumpy. Then again, one will in general think that the least clumpy case
compares to that the occasions are consistently divided throughout the course of events. Among the current
"hot hand" measures, the vast majority of them have these two properties (see Table 1).
We contend that Continuity is a fundamental property. It would "have neither rhyme nor reason" if a minuscule
change in the event of occasions brings about a sensational move in the level of clumpiness. As

CHAPTER-2
Literature Review

an outcome, neither one of the runs like estimates which center around the flow of occasions nor any hard
edge technique fulfills this property.
The Convergence property, while fairly loose, is the critical articulation of individuals'
instinctive idea of clumpiness. Sadly, no current hot hand measures have this property. Take the runs test for instance, it just considers the quantity of runs, and completely
overlooks the data contained by the lengths of runs. However long visits are not continuous,
it is considered not very clumpy despite the fact that every one of the occasions are as of now near one another.
On the off chance that there are just two occasions, it is extremely clear how the Convergence property would function.
Yet, for the instance of north of two occasions, a cautious decision of definition is required for "development of
7
occasion" to ensure it doesn't struggle with different properties. As the event of one
occasion moves towards another visit, it likewise get away from the other neighbour(if any). It is
normal to utilize the negligible worth of the distances to two neighbors as a depiction of nearby
intermingling. The more modest the worth, the more joined it is. Presently the Convergence property
can be reworded as follows: given the wide range of various occasions, the clumpiness measure should move
a similar way with neighborhood union. Subsequently, when one occasion goes from the left
endpoint to right endpoint inside its neighborhood span, the action acts like a U-bend, and
the other way around, which is represented previously. Specifically, the Convergence property could be recorded
as $d(Clumpy)/d(IET) < 0$ when IETs are made more modest.
For a sensible proportion of clumpiness, the Minimum and Maximum property sets its
limits, the Continuity property administers its speed, and the Convergence property indicates
its bearing. These four properties consolidated give a complete depiction of the
measure elements. Prior to presenting our new measures, let us sum up which wanted
properties the "hot hand" measures have. As shown in Table 1, albeit the vast majority of them
fulfill the Minimum and Maximum property, not even one of them have the Continuity or Convergence
property.
The intricacy of examples across a dataset, the intricacy of examples

CHAPTER-2
Literature Review

inside a bunch and the
intricacy of the foundation in which bunches are frequently inserted, make a general measure
of clumpiness hard to accomplish. Albeit the advancement of wanted properties have added
some helpful principles to building new measures, it actually passes on sufficient space for specialists to
make their own actions one case at a time case. Thusly, the current paper won't investigate all
the potential decisions; all things being equal, we just consider an exceptional class by adding the accompanying three
conditions:

• Capacity of between occasion times(IETs)

The conversation of wanted properties prompts us to consider measures built utilizing
IETs, by which we can all the more likely control the action elements. IETs not just give
adaptable reliance structures, yet in addition loosen up the dependence on succession of occasions.
Likewise, the actions utilizing IETs are not confined to the discrete case, and can be
handily reached out to survey constant appearance times. One more decent property of picking a
class of clumpiness measures dependent on IETs is their invariance to advance or in reverse
orderings of the perceptions, or any change that doesn't impact the IETs, or
does as such (given we scale by the reach) in an absolutely corresponding way.

• Balance

Because of the Convergence property that the clumpiness measure should arrive at the base
esteem when the occasion sits in its stretch, we use balance to stay away from a
mutilation.

• Convexity

The Convergence property directs the adjusting course of the action, however what might be said about
the evolving rate? Clumpiness can be considered as an obstruction file notable in
the designing writing. Allow us to envision a test like this, a ball is tied between
9
two holders with two springs of equivalent length. Thus, the steady condition is the ball
winding up in the center. As the ball is pushed toward one holder, it requires more and
more power to follow up on the ball. In particular, the power

develops quadratically. Relating this
similarity to clumpiness, the changing pace of the action ought to be expanding when one
occasion is moving toward each other. All in all, convexity is a sensible supposition,
which takes into consideration a more delicate identification of limits.

CHAPTER-3
Data

In the past 10 years, the applications of artificial neural networks have developed outstandingly from image segmentation to speech recognition. One notably successful application of deep learning is deep embedding, a technique to transform the input data into a more useful representation, a list of real numbers called vectors. During the training on a supervised machine learning prediction task, the parameters of the neural network – the weights – are the embeddings that will modify to minimize the loss. These resulting embedding vectors have closer representation in the embedding space for the inputs from a similar category. Deep embedding is widely used to make recommendations by finding the nearest neighbors in the embedding space.

Every eCommerce application or retailer has millions of products. Identifying similar products can be used for recommendation and search. Our task is to build a product recommendation system from the product image and text description.

ConvNet will classify the product images and the LSTM network with an embedding layer will classify the description text. We have used a supervised learning process to train the model by labeling the samples by their category. The neural network models learn while training by a feedback process called backpropagation. This involves comparing the output produced by the network with the actual output and using the difference between them to modify the weights of the connections between the units in the neural network. We assume that once the model successfully trained the output or features from the final layer are embedded. Being more specific, the output layer will generate a similar dense vector for the samples from similar categories. We can create the embedding space by using these vectors and similarities.

In the beyond 10 years, the utilizations of counterfeit neural organizations have grown extraordinarily from picture division to discourse acknowledgment. One strikingly effective use of profound learning is profound installing, a procedure to change the info information into a more helpful portrayal, a rundown of genuine numbers called vectors. During the preparation on a managed AI expectation task, the boundaries of the neural organization – the loads – are the embeddings that will alter to limit the misfortune. These subsequent inserting vectors have nearer portrayal in the implanting space for the contributions from a comparable class. Profound inserting is broadly used to make suggestions by finding the closest neighbors in the implanting space.

Each eCommerce application or retailer has a great many items. Distinguishing comparable items can be utilized for proposal and search. Our assignment is to fabricate an item proposal framework from the item picture and text depiction.

ConvNet will order the item pictures and the LSTM network with an implanting layer will characterize the depiction text. We have utilized a managed learning interaction to prepare the model by

CHAPTER-3
Data

marking the examples by their classification. The neural organization models learn while preparing by an input cycle called backpropagation. This includes looking at the result delivered by the organization with the genuine result and utilizing the contrast between them to change the loads of the associations between the units in the neural organization. We expect that once the model effectively prepared the result or elements from the last layer are implanted. Being more explicit, the result layer will produce a comparable thick vector for the examples from comparative classes. We can make the installing space by utilizing these vectors and likenesses.

This dataset comprises of in excess of 42000 item data, for example, pictures and short portrayals having a place with 21 classifications. The train envelope contains all the item pictures of size 100X100 used to prepare the model. The test organizer contains pictures to assess model execution. data.csv document contains the accompanying data: -

ImgId: interesting id of the item. Additionally, every one of the pictures are saved by this name.

title: Name of the item - portrayal: a short depiction of the

This dataset consists of more than 42000 product information such as images and short descriptions belonging to 21 categories. The train folder contains all the product images of size 100X100 used to train the model. The test folder contains images to evaluate model performance. data.csv file contains the following information: -

ImgId: unique id of the product. Also, all the images are saved by this name.

title: Name of the product - description: a short description of the product - category: name of the category the product belongs to.

	transaction_id	cust_id	tran_date	year	month	day	prod_subcat_code	prod_cat_code	Qty
0	80712190438	270351	2014-02-28	2014	2	Friday	1	1	-5
1	29258453508	270384	2014-02-27	2014	2	Thursday	5	3	-5
2	51750724947	273420	2014-02-24	2014	2	Monday	6	5	-2
3	93274880719	271509	2014-02-24	2014	2	Monday	11	6	-3
4	51750724947	273420	2014-02-23	2014	2	Sunday	6	5	-2

Table 1

CHAPTER-3
Data

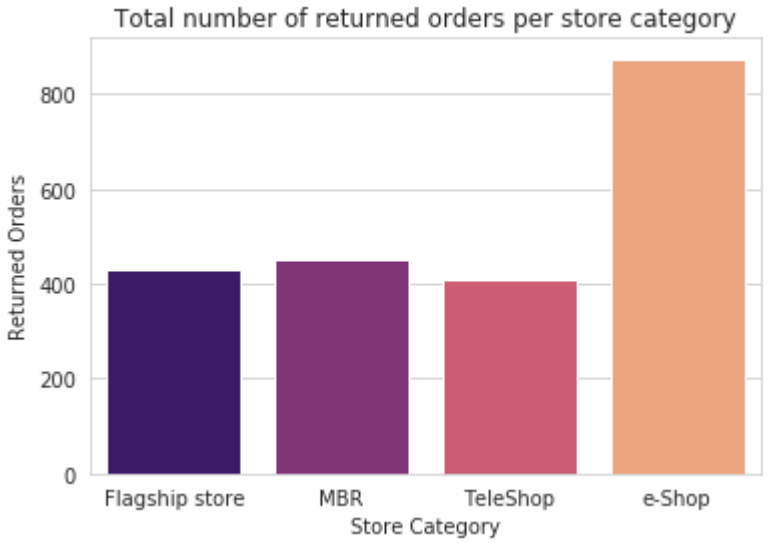

Figure 1

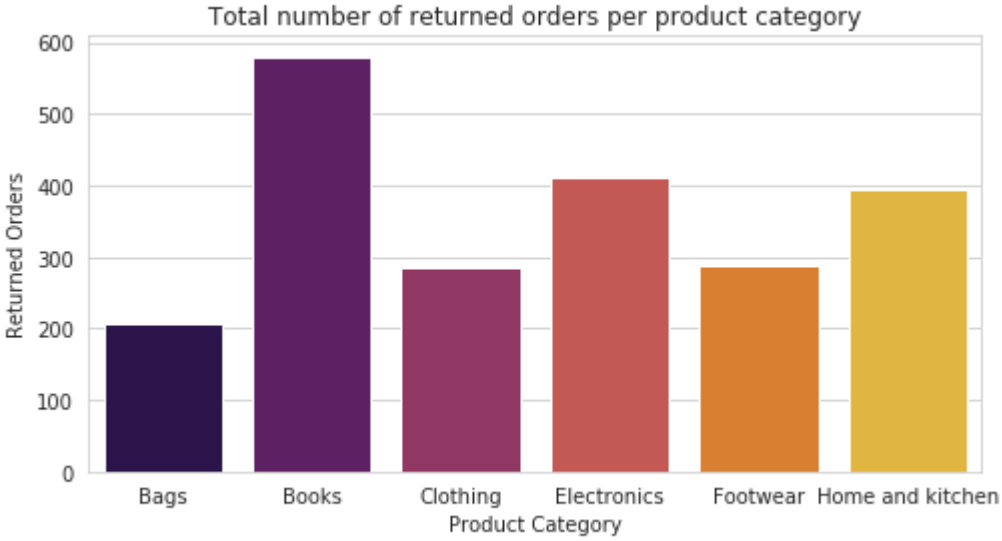

Figure 2

CHAPTER-3
Data

Figure 3

CHAPTER-4
Model Architecture

The following is a step-by-step, do-it-yourself approach to RFM segmentation.

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning, which will be the main focus of this article.

How can an algorithm be represented as a tree?

For this let's consider a very basic example that uses titanic data set for predicting whether a passenger will survive or not. Below model uses 3 features/attributes/columns from the data set, namely sex, age and sibsp (number of spouses or children along).

Image taken from wikipedia

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, you will need to trim it down for it to look beautiful. Lets start with a common technique used for splitting.

CHAPTER-4
Model Architecture

Recursive Binary Splitting

In this procedure all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

Consider the earlier example of tree learned from titanic dataset. In the first split or the root, all attributes/features are considered and the training data is divided into groups based on this split. We have 3 features, so will have 3 candidate splits. Now we will calculate how much accuracy each split will cost us, using a function. The split that costs least is chosen, which in our example is sex of the passenger. This algorithm is recursive in nature as the groups formed can be sub-divided using same strategy. Due to this procedure, this algorithm is also known as the greedy algorithm, as we have an excessive desire of lowering the cost. This makes the root node as best predictor/classifier.

Cost of a split

Lets take a closer look at cost functions used for classification and regression. In both cases the cost functions try to find most homogeneous branches, or branches having groups with similar responses. This makes sense we can be more sure that a test data input will follow a certain path.

Regression : $\text{sum}(y - \text{prediction})^2$

Lets say, we are predicting the price of houses. Now the decision tree will start splitting by considering each feature in training data. The mean of responses of the training data inputs of particular group is considered as prediction for that group. The above function is applied to all data points and cost is calculated for all candidate splits. Again the split with lowest cost is chosen. Another cost function involves reduction of standard deviation, more about it can be found here.

Classification : $G = \text{sum}(pk * (1 - pk))$

A Gini score gives an idea of how good a split is by how mixed the response classes are in the groups created by the split. Here, pk is proportion of same class inputs present in a particular group. A perfect class purity occurs when a group contains all inputs from the same class, in which case pk is either 1 or 0 and $G = 0$, where as a node having a 50–50 split of classes in a group has the worst purity, so for a binary classification it will have $pk = 0.5$ and $G = 0.5$.

CHAPTER-4
Model Architecture

When to stop splitting?

You might ask when to stop growing a tree? As a problem usually has a large set of features, it results in large number of split, which in turn gives a huge tree. Such trees are complex and can lead to overfitting. So, we need to know when to stop? One way of doing this is to set a minimum number of training inputs to use on each leaf. For example we can use a minimum of 10 passengers to reach a decision(died or survived), and ignore any leaf that takes less than 10 passengers. Another way is to set maximum depth of your model. Maximum depth refers to the the length of the longest path from a root to a leaf.

Pruning

The performance of a tree can be further increased by pruning. It involves removing the branches that make use of features having low importance. This way, we reduce the complexity of tree, and thus increasing its predictive power by reducing overfitting.

Pruning can start at either root or the leaves. The simplest method of pruning starts at leaves and removes each node with most popular class in that leaf, this change is kept if it doesn't deteriorate accuracy. Its also called reduced error pruning. More sophisticated pruning methods can be used such as cost complexity pruning where a learning parameter (alpha) is used to weigh whether nodes can be removed based on the size of the sub-tree. This is also known as weakest link pruning.

Advantages of CART

Simple to understand, interpret, visualize.

Decision trees implicitly perform variable screening or feature selection.

Can handle both numerical and categorical data. Can also handle multi-output problems.

Decision trees require relatively little effort from users for data preparation.

Nonlinear relationships between parameters do not affect tree performance.

Disadvantages of CART

CHAPTER-4
Model Architecture

Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.

Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.

Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.

Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

This is all the basic, to get you at par with decision tree learning. An improvement over decision tree learning is made using technique of boosting. A popular library for implementing these algorithms is Scikit-Learn. It has a wonderful api that can get your model up an running with just a few lines of code in python.
Note that with the aid of software, RFM segmentation – as well as other, more sophisticated types of segmentation – can be done automatically, with more accurate results.
Step 1
The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer. The raw data for doing this, which should be readily available in the company's CRM or transactional databases, can be compiled in an Excel spreadsheet or database:

- Recency is simply the amount of time since the customer's most recent transaction (most businesses use days, though for others it might make sense to use months, weeks or even hours instead).
- Frequency is the total number of transactions made by the customer (during a defined period).
- Monetary is the total amount that the customer has spent across all transactions (during a defined period).

Step 2
The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M), using Excel or another tool. Unless using specialized software, it's recommended to divide the customers into four tiers for each dimension, such that each customer will be assigned to one tier in each dimension:

Recency	**Frequency**	**Monetary**
R-Tier-1 (most recent)	F-Tier-1 (most frequent)	M-Tier-1 (highest spend)
R-Tier-2	F-Tier-2	M-Tier-2

CHAPTER-4
Model Architecture

R-Tier-3	F-Tier-3	M-Tier-3
R-Tier-4 (least recent)	F-Tier-4 (only one transaction)	M-Tier-4 (lowest spend)

Table 2

This results in 64 distinct customer segments (4x4x4), into which customers will be segmented. Three tiers can also be used (resulting in 27 segments); using more than four, however, is not recommended (because the difficulty in use outweighs the small benefit gain from the extra granularity).

As mentioned above, more sophisticated and less manual approaches – such as k-means cluster analysis – can be performed by software, resulting in groups of customers with more homogeneous characteristics.

Step 3

The third step is to select groups of customers to whom specific types of communications will be sent, based on the RFM segments in which they appear.

It is helpful to assign names to segments of interest. Here are just a few examples to illustrate:

- **Best Customers** – This group consists of those customers who are found in R-Tier-1, F-Tier-1 and M-Tier-1, meaning that they transacted recently, do so often and spend more than other customers. A shortened notation for this segment is 1-1-1; we'll use this notation going forward.
- **High-spending New Customers** – This group consists of those customers in 1-4-1 and 1-4-2. These are customers who transacted only once, but very recently and they spent a lot.
- **Lowest-Spending Active Loyal Customers** – This group consists of those customers in segments 1-1-3 and 1-1-4 (they transacted recently and do so often, but spend the least).
- **Churned Best Customers** – This segment consists of those customers in groups 4-1-1, 4-1-2, 4-2-1 and 4-2-2 (they transacted frequently and spent a lot, but it's been a long time since they've transacted).

Marketers should assemble groups of customers most relevant for their particular business objectives and retention goals.

Step 4

The fourth step actually goes beyond the RFM segmentation itself: crafting specific messaging that is tailored for each customer group. By focusing on the behavioral patterns of particular groups, RFM marketing allows marketers to communicate with customers in a much more effective manner.

Again, here are just some examples for illustration, using the groups we named above:

- **Best Customers** – Communications with this group should make them feel valued and appreciated. These customers likely generate a

disproportionately high percentage of overall revenues and thus focusing on keeping them happy should be a top priority. Further analyzing their individual preferences and affinities will provide additional opportunities for even more personalized messaging.

- **High-spending New Customers** – It is always a good idea to carefully "incubate" all new customers, but because these new customers spent a lot on their first purchase, it's even more important. Like with the Best Customers group, it's important to make them feel valued and appreciated – and to give them terrific incentives to continue interacting with the brand.
- **Lowest-Spending Active Loyal Customers** – These repeat customers are active and loyal, but they are low spenders. Marketers should create campaigns for this group that make them feel valued, and incentivize them to increase their spend levels. As loyal customers, it often also pays to reward them with special offers if they spread the word about the brand to their friends, e.g., via social networks.
- **Churned Best Customers** – These are valuable customers who stopped transacting a long time ago. While it's often challenging to re-engage churned customers, the high value of these customers makes it worthwhile trying. Like with the Best Customers group, it's important to communicate with them on the basis of their specific preferences, as known from earlier transaction data.

Of course, deciding which groups of customers to target and how to best communicate with them is where the art of marketing comes in!

What is "clumpiness" in customer data, and why it matters:

One of the most established practices in the field of marketing and customer valuation is to summarize a customer using what's called RFM segmentation — recency, frequency, monetary value — which means I take everything I know about my customer and I compute just three simple numbers: How recently did they buy? How frequently do they buy? And when they buy, how much money do they spend? … It's the basis on which most companies decide who are the valuable customers and who are the non-valuable customers.

My research … says that's not a complete characterization of customers. You have to add one more letter to RFM, and I call that "C," which [stands for] clumpiness Twitter , which means some customers do buy in a regular pattern. Historically, if you bought orange juice, if you bought diapers, you bought things in a regular pattern. But clumpiness refers to the fact that people buy in bursts.

CHAPTER-4
Model Architecture

And those burst periods indicate something very different about the customer and that those customers could be extremely valuable.

On the key takeaways:

The key takeaways of my research are very simple. Let's imagine you want … to predict who are going to be the valuable customers in the future. And you have four things you can use to predict it. As I mentioned: recency, frequency, monetary value and let's say the marketing spend towards the customer. Those are the classic ways in which companies build what are called scoring models. I'm claiming you need to add one more number, and that's C — how clumpy the customer is. This is no more difficult to compute than R, F and M. You can do it in Excel. It's very quick to compute. You can compute it for literally 100 million customers in a second.

Burst periods indicate something very different about the customer and that those customers could be extremely valuable.

And the findings of my research suggest that higher clumpy customers are worth more out of sample, meaning in their future value, even after controlling for RFM and marketing expenditure — which means we have found another variable that firms should track [concerning] our customers and use it to predict their worth in the future.

The most surprising conclusions:

Two things surprised me about my conclusions. One is I just figured that this RFM based segmentation, which had been around for so long and is used by so many firms, had been validated in the sense that there wasn't anything else simple out there that could help explain customer value. You can do all kinds of fancy web-scraping and all kinds of other variable construction, but clumpiness is so simple.

CHAPTER-4
Model Architecture

So, first, I was surprised that that had been missed — that, in other words, hot and cold periods are indicative of something about the customer. I think the second part that surprised me, at least in the data sets I've analyzed, [is that] it's true for digital and online consumption goods, but it's not true for regular consumer package goods. In other words, with historical models I can see why they fit fine, because you buy toilet paper in a regular pattern; you buy orange juice in a regular pattern. But you don't consume Hulu in a regular pattern. You don't bid on auctions at eBay in a regular pattern. You don't buy books at Amazon on a regular pattern.

If you look at historically purchased goods, clumpiness really isn't there. But if you look in the new wave, the new economy, clumpiness is pervasive in every data set I've analyzed.

KNOWLEDGE@WHARTON HIGH SCHOOL

On the practical implications:

I think of all the research I've done over my … 20-year career, this is probably the most practical thing I've done. The work I do tends to be what I call fancy, complex statistical modeling. And this isn't about statistical modeling. This is about a number — clumpiness — that firms can actually compute today. They don't need to collect any additional data. It's the same data they're using to compute R, F and M and customer lifetime value. And they can figure out how much value it adds to predicting customer value. Your rank ordering of customers will change. Your decisions about which customers are valuable to reactivate — imagine customers have churned — well, which ones are valuable to reactivate?

My claim is the clumpy ones, even though they've churned … are the ones to reactivate. If you reactivate them, they'll come back and be clumpy again, and do a lot of stuff in the future. So, I think it has huge practical value. And the beauty of it is, if you go to my website I have an Excel sheet there that has worked out examples. It actually

CHAPTER-4
Model Architecture

has an Excel sheet that you can just download and you can start using clumpiness today.

What new rules, procedural changes or strategies would you suggest as a result of your research?

A lot of people today talk about big data. I love big data, but I'll tell you what I love even more than big data. I love data compression. And what I mean by data compression is, you can collect thousands and thousands of variables on people now. You can track where they are and you can track what they bought, what web pages they looked at. But that's not science. That's data collection. Now a question is, which of that information is actually useful for the business problem at hand? And that's what I call data compression.

So, the way I view clumpiness is as an addition to traditional variables like RFM, marketing activity and stuff like that — I view it as a form of, let's call it increased data compression. I'm just telling you that you need to keep a little bit more data. You can't compress things down to three numbers. You've got to compress it down to four.

I've done a lot of work on clumpiness, I know it exists across industries. I know it can be of predictive value. Here's what I don't know: what causes it…. I've related marketing activity to clumpiness. Firms can try to make you clumpy by sending you an e-mail, by sending you a catalog, by targeting you, etc.

But I haven't really studied yet what's the optimal way in which firms should target you, knowing that clumpiness exists. I haven't looked at, for example, do you consume more clumpy content if it's a series? Imagine watching "Breaking Bad" or "Mad Men" or something like that. Or imagine you're a firm and you're trying to sell a suite of products like a facial care line and a moisturizer line, and all this other stuff. Should you package it together and make it seem like people are progressing towards a goal?

CHAPTER-4
Model Architecture

If you look at historically purchased goods, clumpiness really isn't there. But if you look in the new wave, the new economy, clumpiness is pervasive in every data set I've analyzed.

So … I know mathematically how to compute it. I know it's trivial for firms to do. I know it's predictive. But the part that's left unknown to me is the psychology of why, which is why I'm partnering right now with a lot of my more consumer psychology-oriented colleagues. We're going to start running a lot of behavioral experiments in the lab to try to get to the underlying psychology of why people behave in a clumpy fashion.

On "Clumpy" vs. "Bingeing"

I like the word "clumpiness." Other people like the word "bingeing." The reason I like clumpiness is that it refers to the opposite, which is non-clumpy, which is kind of equally spaced arrivals or equally spaced purchases. I don't think I've seen a story about clumpiness, but any time you see a story about people bingeing [on] content or people consuming things – "a student sat up for 18 hours watching this" — it applies. And the concept is so pervasive: Every time I talk to managers, students or academics about it, everyone believes it exists.

Dispelled misperceptions:

My research dispels the idea that in some sense customers can just be categorized by a simple set of numbers. You need to go a little bit beyond that. You need to go a little bit beyond what I would call simple theories of how people behave. If you look at recency, frequency, monetary value, which is kind of the historical basis of consumer behavior, it basically ignores what I call the inter-arrival times. It basically says, I can take all the data — like it was a two-day window and then a four-day window and then a three-day window, then a six-day window — I can throw all of that away and all I need

to know is when's the last time you came and how many times had you come?"

What this dispels is [the notion] that the arrival pattern of people is uninformative. It's very informative. People [who] come in bursts, then go away and then come back in bursts and then go away … those people are fundamentally different. I personally believe there are clumpy-type people and non-clumpy-type people.

What we've also shown is, it varies by product categories. So, we've found, for example, that women tend to be more clumpy than men. We've found that younger people tend to be more clumpy in their consumption than older people. So, I think [one of] the myths that we're going to dispel is that not only are all people created equal, but that there are simple ways to just categorize all people into a certain type.

How the study stands apart:

There's a whole class of mathematical models that have been popularized — although they've been around for 50 years — over the last 10 years called "hidden Markov models." … Let's imagine there are two states of the world. You're in a hot state or a cold state, and you rotate back and forth between a hot and a cold state. That mathematical model is clumpiness. You're hot, you do a lot of stuff. You're cold, you don't. Hot, cold, hot, cold…. What I wanted to do was to bring to the practitioner a way that they could compute a simple number. It's a statistic. It's not a statistics paper. It's a paper about a number, a statistic, as we call it.

You just compute the number and then do what you want with it. You could try to use it to predict customer value. You could use it to see whether men are more clumpy than women. You could use it to segment people. That's what typifies and separates this work — it's a simple metric-based approach that practitioners can use. It's not a fancy modeling based approach. But they're both trying to cover the same problem.

CHAPTER-4
Model Architecture

The work I'm doing isn't ivory tower mathematics. It's a simple number that someone can compute.

An example of how "clumpiness" might be used:

What we've studied so far with reaching clumpy customers is whether e-mail, catalogs, different types of marketing channels are more effective. What we found, not surprisingly, is e-mail has more of a short-term effect, as you would expect. [A] catalog has more of a longer-term effect.

What we've yet to really understand is — are there certain words in an e-mail or a catalog or a video campaign that will engage or, if you'd like, cause people to be more clumpy? Are there certain topics that are more clumpy, some product categories that will necessarily be more clumpy? All we've done so far is to establish that the phenomenon exists. I know it exists across lots of industries. I know certain types of people tend to be more clumpy.

The part that I haven't done, which is shocking because I'm a professor of marketing, is talk about the marketing implications of it yet. That's going to require bigger and newer data sets that allow me to link things about marketing campaigns to people's clumpy behavior. I know how to do it. I just need richer and better data to do it.

What's next?

I'm thinking of about three different streams to follow up this research. First of all, I'd be thrilled to just analyze more data sets and prove how pervasive the clumpiness measure is. I've analyzed data sets from Amazon, from CDNow, eBay, Hulu, YouTube and also from some traditional consumer package goods companies. [Now,] I just want to apply it to new data sets.

The second is I want to understand the psychological processes. Why are people behaving in a clumpy fashion? The third and final piece is

CHAPTER-4
Model Architecture

I want to relate marketing activity to clumpiness. Now, that's going to require not just people's behaviors — like what did they do? What web sites did they visit? What did they purchase? [It will also require] information about the marketing campaigns themselves, possibly even the copy of the marketing campaign, which channels they were sent through — and that's going to allow me to come up with optimization — ways for firms to optimize their marketing campaigns to activate clumpiness

• Agitated Best Customers – This section comprises of those clients in bunches 4-1-1, 4-1-2, 4-2-1 and 4-2-2 (they executed oftentimes and spent a great deal, however it's been quite a while since they've executed).

Advertisers ought to gather gatherings of clients generally pertinent for their specific business destinations and maintenance objectives.

Stage 4

The fourth step really goes past the RFM division itself: making explicit informing that is custom-made for every client bunch. By zeroing in on the standards of conduct of specific gatherings, RFM advertising permits advertisers to speak with clients in a substantially more powerful way.

Once more, here are only a few models for delineation, utilizing the gatherings we named previously:

• Best Customers – Communications with this gathering should cause them to feel esteemed and appreciated. These clients probably produce an excessively high level of generally incomes and accordingly zeroing in on keeping them cheerful ought to be a main concern. Further examining their singular inclinations and affinities will give extra freedoms to much more customized informing.

• High-spending New Customers – It is consistently smart to painstakingly "brood" every single new client, but since these new clients spent a ton on their first buy, it's considerably more significant. Like with the Best Customers bunch, it's critical to cause them to feel esteemed and appreciated – and to give them spectacular motivators to keep connecting with the brand.

• Least Spending Active Loyal Customers – These recurrent clients are dynamic and faithful, yet they are low spenders. Advertisers ought to make lobbies for this gathering that cause them to feel esteemed, and boost them to expand their spend levels. As

steadfast clients, it regularly additionally pays to compensate them with unique offers assuming that they spread the word about the brand to their companions, e.g., through informal organizations.

• Beaten Best Customers – These are important clients who quit executing quite a while in the past. While it's frequently difficult to reconnect beaten clients, the high worth of these clients makes it advantageous difficult. Like with the Best Customers bunch, it's critical to speak with them based on their particular inclinations, as known from prior exchange information.

Obviously, concluding which gatherings of clients to target and how to best speak with them is the place where the specialty of promoting comes in!

The Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).

Visualization of a Random Forest Model Making a Prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

There needs to be some actual signal in our features so that models built using those features do better than random guessing.

CHAPTER-4
Model Architecture

The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

An Example of Why Uncorrelated Outcomes are So Great

The wonderful effects of having many uncorrelated models is such a critical concept that I want to show you an example to help it really sink in. Imagine that we are playing the following game:

I use a uniformly distributed random number generator to produce a number.

If the number I generate is greater than or equal to 40, you win (so you have a 60% chance of victory) and I pay you some money. If it is below 40, I win and you pay me the same amount.

Now I offer you the the following choices. We can either:

Game 1 — play 100 times, betting $1 each time.

Game 2— play 10 times, betting $10 each time.

Game 3— play one time, betting $100.

Which would you pick? The expected value of each game is the same:

Expected Value Game 1 = (0.60*1 + 0.40*-1)*100 = 20

Expected Value Game 2= (0.60*10 + 0.40*-10)*10 = 20

Expected Value Game 3= 0.60*100 + 0.40*-100 = 20

Outcome Distribution of 10,000 Simulations for each Game

What about the distributions? Let's visualize the results with a Monte Carlo simulation (we will run 10,000 simulations of each game type; for example, we will simulate 10,000 times the 100 plays of Game 1). Take a look at the chart on the left — now which game would you pick? Even though the expected values are the same, the outcome distributions are vastly different going from positive and narrow (blue) to binary (pink).

Game 1 (where we play 100 times) offers up the best chance of making some money — out of the 10,000 simulations that I ran, you make money in 97% of them! For Game 2 (where we play 10 times) you make money in 63% of the simulations, a drastic decline (and a drastic increase in your probability of losing money). And Game 3 that we only play once, you make money in 60% of the simulations, as expected.

CHAPTER-4
Model Architecture

Probability of Making Money for Each Game

So even though the games share the same expected value, their outcome distributions are completely different. The more we split up our $100 bet into different plays, the more confident we can be that we will make money. As mentioned previously, this works because each play is independent of the other ones.

Random forest is the same — each tree is like one play in our game earlier. We just saw how our chances of making money increased the more times we played. Similarly, with a random forest model, our chances of making correct predictions increase with the number of uncorrelated trees in our model.

If you would like to run the code for simulating the game yourself you can find it on my GitHub here.

Ensuring that the Models Diversify Each Other

So how does random forest ensure that the behavior of each individual tree is not too correlated with the behavior of any of the other trees in the model? It uses the following two methods:

Bagging (Bootstrap Aggregation) — Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging.

Notice that with bagging we are not subsetting the training data into smaller chunks and training each tree on a different chunk. Rather, if we have a sample of size N, we are still feeding each tree a training set of size N (unless specified otherwise). But instead of the original training data, we take a random sample of size N with replacement. For example, if our training data was [1, 2, 3, 4, 5, 6] then we might give one of our trees the following list [1, 2, 2, 3, 6, 6]. Notice that both lists are of length six and that “2” and “6” are both repeated in the randomly selected training data we give to our tree (because we sample with replacement).

Node splitting in a random forest model is based on a random subset of features for each tree.

Feature Randomness — In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a

random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

Let's go through a visual example — in the picture above, the traditional decision tree (in blue) can select from all four features when deciding how to split the node. It decides to go with Feature 1 (black and underlined) as it splits the data into groups that are as separated as possible.

Now let's take a look at our random forest. We will just examine two of the forest's trees in this example. When we check out random forest Tree 1, we find that it it can only consider Features 2 and 3 (selected randomly) for its node splitting decision. We know from our traditional decision tree (in blue) that Feature 1 is the best feature for splitting, but Tree 1 cannot see Feature 1 so it is forced to go with Feature 2 (black and underlined). Tree 2, on the other hand, can only see Features 1 and 3 so it is able to pick Feature 1.

So in our random forest, we end up with trees that are not only trained on different sets of data (thanks to bagging) but also use different features to make decisions.

And that, my dear reader, creates uncorrelated trees that buffer and protect each other from their errors.

Conclusion

Random forests are a personal favorite of mine. Coming from the world of finance and investments, the holy grail was always to build a bunch of uncorrelated models, each with a positive expected return, and then put them together in a portfolio to earn massive alpha (alpha = market beating returns). Much easier said than done!

Random forest is the data science equivalent of that. Let's review one last time. What's a random forest classifier?

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

What do we need in order for our random forest to make accurate class predictions?

We need features that have at least some predictive power. After all, if we put garbage in then we will get garbage out.

The trees of the forest and more importantly their predictions need to be uncorrelated (or at least have low correlations with each other). While the algorithm itself via feature randomness tries to engineer these low

correlations for us, the features we select and the hyper-parameters we choose will impact the ultimate correlations as well.