Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics

Exploring Random Forest Models for the Channel Attribution problem

Govert Verbakel (544180)

Supervisor:	K. Gruber
Second assessor:	M. Mueller
Date final version:	1st July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

Abstract

This research paper proposes the use of a random forest model to solve the channel attribution problem, aiming to overcome the limitations of the widely-used Markov model. The Markov model is constructed through simulating user paths and assessing them by assigning attribution levels. The study uses SHAP values to explain predictions done by the random forest model. These SHAP values are interpreted as Shapley attribution levels, as they explain how a feature (channel) contributes to a prediction (conversion). The applicability of the random forest is evaluated by comparing its Shapley attribution levels with those of the Markov model and a benchmark model. However, the findings reveal that globally explaining the feature importance of the random forest model using the SHAP values, provides limited insight into the actual contribution of features to predictions. Overall, the random forest model is not suitable for solving the channel attribution problem since it lacks interpretability and fails to identify the driving channels for conversions.

1 Introduction

When the internet made its first entrance in the 1960's for researchers to share information, nobody would believe that 50 years later this digital communication channel would become a trillion-dollar industry. The rise of the internet enables retailers to reach new customers, maintain existing customers and most importantly it enables marketers to introduce personal advertising and promotions. Since social media has also made a big impact on the trends in online marketing, marketers need to know how to spend the marketing dollars effectively and how to promote the right content to the right customer at the right time. To be able to optimally allocate their budget, the marketers require to know what works and what does not. This is where the term attribution comes in place. Attribution is the central question in online advertising. So much even that the Marketing Science Institute (MSI) has proposed attribution as the single most important research priority over the past several years (Institute, 2016, 2017, 2018). Channel attribution refers to the process of assigning credit to different marketing channels that contribute to a conversion. Social media, e-mail and an online display advertisement are all examples of such a marketing channel. Having an understanding of attribution as a marketer, helps understanding the effectiveness of marketing efforts, this allows for optimal resource allocation and marketing strategies. Making accurate and substantiated estimates of the attribution per channel, all starts with having a well thought off underlying model.

A well thought off underlying model is necessary to map a customer journey. A customer journey is a path that a customer follows before deciding to buy the product. Every step in this path is an interaction of the customer with a traffic source, for example social media. These traffic sources are called touch-points. One such underlying model is the Markov model, which holds a huge potential in capturing the dynamic nature of a customer's behaviour. The idea of the Markov model originated from the concept of latent states, which represent milestones in a customer's decision-making process. A popular framework that represents these latent states, is the AIDA model, where the latent states are Attention, Interest, Desire and Action. They explain what psychological states a customer goes through on their way to potential conversion. Previously Singal, Besbes, Desir, Goyal and Iyengar (2019) have indeed shown that a Markov model is an excellent option to structure the customer's journey. It provides a probabilistic framework to analyze customer journeys and attribute conversions to specific channels. To model the customer journey, the channels are considered as states of the Markov Chain, meaning the Markov Chain represents the customer travelling from state (channel) to state (channel). Customer behavior can be extremely uncertain and random, which the probabilistic nature of a Markov model can capture. However, the Markov model has its limitations. First of all the Markov property assumes that the future state depends only on the current state and is independent of the history of previous states. Considering the customer's decision making process, a customer's decision does certainly not only depend on the last channel the customer has interacted with. Again referring to the AIDA, it is a long process of creating awareness. Secondly a Markov model is limited by the number of channels (states) it can work with. When including more and more channels, the dimensions of the Markov model become too high, and the problem becomes infeasible. With these limitations in mind we therefore propose to solve this attribution problem using a random forest model. A Random Forest model is a Machine Learning approach which has no problem with high-dimensional data and uses all information to make predictions. To research whether Random Forest model is a good alternative both models need to be evaluated. As mentioned, the Markov model provides a probabilistic framework to analyze customer journeys. From a dataset transitioning probabilities between different marketing channels are obtained. These transition probabilities are used to simulate customer paths. Secondly to evaluate the performance of a Random Forest as a predictive model for the online attribution problem, we use 70% of the existing data as training data. The other 30%, of the data is used as testing data to predict conversion rates. A path, more specifically, the frequency of each channel in a path, is used as input. After building both predictive models, we evaluate them by calculating the attribution levels for the channels. The Shapley value is used as a method to quantify the contribution of each channel to a conversion. It is a commonly used method to approaching the attribution because theoretically it attributes value fairly. The Shapley value is so well applicable because it allows to explain predictive models, through feature importance of every variable. Finally two different heuristics are used to obtain attribution levels per channel, namely the Last-Touch Attribution (LTA) and the Uniform attribution.

The main finding is that the random forest model is not a good underlying model to solve the channel attribution problem. In order for the random forest to predict reasonable conversion predictions for a path, SMOTE sampling is necessary. The SHAP values, which represent the feature importance in a prediction, are used as a global explainer for the predictions done by the random forest model. They are interpreted as the Shapley attribution level for the individual channels. However, the Shapley values, determined by the SHAP values, did not show good results when compared to the Shapley values of the Benchmark model. The Markov model showed to be a suitable underlying model for the channel attribution problem. The Shapley attribution levels are approximated for the simulated paths and these showed to be very similar to the Shapley attribution levels of the benchmark model.

With this research paper we aimed to show that a random forest model is a suitable underlying model for the channel attribution problem. We also aimed to show the use of explainable AI (XAI) in the channel attribution problem. The applicability of the random forest is evaluated by interpreting the SHAP values as Shapley attribution levels. However comparing the Shapley attribution levels of the random forest model with the levels of the Markov and benchmark model, the random forest model does not show to be a suitable underlying model for the channel attribution problem.

2 Literature review

In this paper we aim to find the best model for structuring customer journeys. In this section we discuss existing literature available on finding the best model for this problem in the online advertising attribution problem. We start by discussing existing literature on probabilistic models, under which the Markov model would classify. Thereafter existing literature on existing machine learning approaches are discussed.

2.1 Probabilistic models

The idea of using a probabilistic model to map the customer's journey originates from the highly dynamic nature of customer's behavior. A customer generally funnels towards conversion through four latent states, namely Attention, Interest, Desire and Action (AIDA). Users move through these stages on their way to conversion. The idea of using a probabilistic model, specifically a Markov model, as an underlying model in marketing was introduced to understand and predict the customer behavior, namely the likelihood of customers moving from one stage of the buying process to the other. The AIDA model focuses on the psychological journey of the customer through different stages, the Markov model tries to analyse and predict the customer's behavior. Markov models can be used to estimate transition probabilities between the stages of the AIDA model. This allows for data-driven insights into the effectiveness of marketing efforts. As stated in the paper by Abhishek, Fader and Hosanagar (2012), the use of heuristic models as attribution schemes is not data-driven, and this is generally considered a shortcoming. There have been several studies that offer more data-driven approaches to the attribution explanation to be able to overcome the weaknesses of Heuristic models. Yadagiri, Saini and Sinha (2015) and Nisar and Yeung (2015) use the Shapley value to explain attribution in their game theory based model.

There are many works available that use a Markov model as an approach to model customer journeys. Previously Singal et al. (2019) have shown a Markov model to identify the structural relations in the journey of the customers through the data and different channels. The researchers use several attribution schemes, they use heuristics i.e. the Last-Touch Attribution (LTA), Incremental Value Heuristic (IVH) and Uniform Attribution. Besides the heuristics, they also approach the channel attribution problem using the Shapley Value, stemming from the cooperative Game Theory. Based on the Shapley value they propose an axiomatic framework under the name Counterfactually-Adjusted Shapley Value (CASV), which is supposed to overcome the shortcomings of the usual Shapley Value in the online advertising context. The CASV is supposed to form a generalizable framework that is widely applicable. Although it is a great approach to explain the attribution better, it is unfortunately not so generalizable. The CASV relies on data where advertiser actions are known. This is unfortunately not very common in channel attribution datasets. Besides not typically observing advertising actions, the counterfactual remains unobserved as well in most cases. The methods proposed in the paper by Singal et al. (2019) are thus difficult to reproduce. This is because the ideas they have in the paper rely heavily on a specific non-generalizable dataset.

2.2 Machine Learning models

Although probabilistic models show the ability to predict, this is not what they are initially designed to do. Machine Learning (ML) models however, are designed to be pure predictive models. Traditional approaches, such as the Markov model, have been extensively used for this purpose. However, with the increased popularity in ML models, the analysis of customer behavior has significantly changed. A paper by Song, Mitnitski, Cox and Rockwood (2004) shows a comparison of machine learning techniques with classical statistical models in health outcomes. The authors show that ML models consistently outperform classical statistical models. Yang, Dyer and Wang (2020) have already shown in their research that machine learning offers accurate and interpretable results when researching attribution in online advertising. The researchers use an interpretable deep learning structure called DeepMTA, this model combines deep learning with an additive feature explanation model, this makes sure we get interpretable online multi-touch attribution. "Evaluation on a real dataset shows the proposed conversion prediction model achieves 91% accuracy.". Furthermore, it is known that the Markov model is limited to a number of states. This means that if we increase the number of states, in our case channels, the problem will become infeasible. ML models on the other hand, show to do an excellent job at working with high-dimensional data (Janitza, Celik & Boulesteix, 2018).

Although there are many advantages when using ML models to perform predictive tasks, there is one main problem. ML models are known to be black box models, meaning the predictions are very hard to explain. In a paper by Chuang et al. (2023), the researchers describe the need for Explainable Artificial Intelligence (XAI) algorithms. In a paper by Balkanski and Singer (2015) the idea of using Shapley values for interpretation of fairness is researched. Merrick and Taly (2020) present a general game formulation for the popular SHAP algorithm. SHAP values are used to calculate feature importance of variables, thus what individual variables contribute to the final prediction. The contribution a variable has to a final prediction can be translated to the attribution levels of a channel to a conversion, determined by the Shapley value.

3 Data

In Singal et al. (2019) the researchers evaluate their proposed methods by implementing their framework on a large-scale real-world data set. Their data set consists of several million user paths, with a few hundred thousand conversion. The conversion data is for a single product sold by a Fortune500 company. To replicate the ideas and methods of the researchers an alternative data set is found and used. In this section we will discuss the data set we use in our research and compare it to the data set used in (Singal et al., 2019).

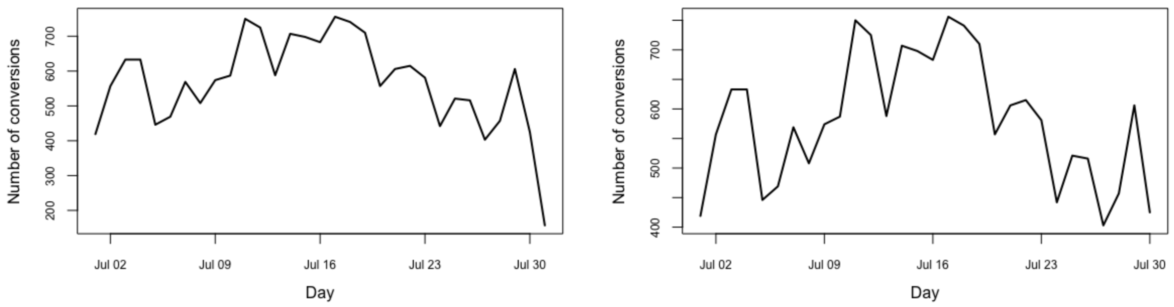
The data set we use to perform our research can be downloaded from the website Kaggle[1]. The data set contains 586,737 marketing touch-points during one month (July 2018). The touch-points belong to 240,108 individual customers with 17,639 conversions. The data set contains six different columns that can be seen in Table 1.

Table 2 shows some summary statistics for our data. It shows the total number of conversions,

Column	Meaning
Cookie	An anonymous customer id, to check the progression of the customer.
Timestamp	Date and time where the visit took place.
Interaction	The type of interaction that took place on the visit.
Conversion	Boolean variable indicating whether a conversion took place.
Conversion value	Revenue created by the potential conversion.
Channel	The marketing channel through which the customer visits the site.

Table 1: Column name with its respective meaning.

[1] See more on: https://www.kaggle.com/code/hughhuyton/multitouch-attribution-modelling/notebook



(a) Daily number of conversions from July 1st until July 31st.

(b) Daily number of conversions from July 1st until July 30th.

Figure 1: The daily number of conversions 1a shows the original data, and Figure 1b shows the cleaned data.

the total conversion rate, the total value of conversions and the average conversion value[2].

Total conversions	17,639
Total conversion rate	3%
Total value of conversions	$110231
Average conversion value	$6

Table 2: Summary statistics including total conversions, average conversion rate, total conversion value and average conversion value.

3.1 Data Cleaning

To gain some insight into our dataset a plot was made that showed the number of conversions per day, see Figure 1a. In Figure 1a a steep decline can be seen from July 30th onwards. The cause of this relatively low observation is unknown, it could for example be that the website malfunctioned. As we are working with predictive models in this paper, a stationary dataset is required, we therefore choose to delete the observations of July 31st. Figure 1b shows the new plot. A more stationary time series can be observed. The daily number of conversions now range from 400 to 756 on the 17th of July.

Furthermore in our dataset there are five possible channels through which the customer funnels, these are Facebook, Instagram, Online Display, Online Video and Paid Search. Comparing this to the data set used in the paper by Singal et al. (2019) there are four different channels namely No-ad, E-mail, Display and Paid Search Click. To compare the results more easily we propose to cluster two channels of the downloaded data set so an equal number of channels is obtained. The channels Facebook and Instagram, now together form the channel Social Media. Because of this clustering we do expect the attribution to be higher for channel Social Media, because the occurrence is counted double. This should be taken into account.

[2] This data can be considered relatively common considering “The average conversion rate across all fourteen industries is 2.9%.”. See more on: https://www.ruleranalytics.com/blog/insight/conversion-rate-by-industry/

4 Methodology

In this paper, the channel attribution in a multi-touch attribution problem is evaluated, using two underlying models and several attribution schemes. This section describes the two distinct models that map the customer journeys and predict the conversion rate. These two models are the Markov model and the Random Forest model. The Markov model makes use of a probabilistic approach to understand how a customer moves towards conversion or leaves the system. The Random Forest Model is a machine learning model that predicts a conversion probability for a path, based on individual decision trees. Furthermore, to evaluate the models, different channel attribution schemes are used, these show how much a single channel contributed to a customer converting. For the Markov model, attribution levels are obtained using the Shapley value, a concept stemming from game theory. The Shapley attribution levels for the random forest model are derived from the SHAP value, a feature importance measure based on the Shapley value. In addition, three different heuristic models are discussed, namely the Last-Touch Attribution (LTA), the Incremental Value Heuristic (IVH) and the Uniform Attribution. The attribution levels obtained from the two models will be evaluated with the attribution levels obtained for the benchmark model. The performance of the predictive models is evaluated with benchmark channel attributions of the model with actual number of conversions.

4.1 Markov Model

The Markov model is a commonly used approach to solving the channel attribution problem. The model is used to map the customers' journey. The model represents a customers' journey towards potential conversion as a sequence of states, in our context, channels. The customer transitions from one state to another over time, where the transitioning depends solely on the state (channel) the customer is currently in. This is the Markov property, where the transitioning to another channel does not depend on the history of previous channels. Using transitioning probabilities, which represent the probabilities of moving from one state (channel) to the other, we can represent the customers journey as a Markov Chain. The transition matrix containing the transition probabilities is estimated using data. The observed frequency of channel transitions allows us to calculate transition probabilities. The applicability of the Markov model as an underlying model is evaluated by simulating customer journeys, which end up in a customer converting or not. The paths that are simulated represent the actual Markov model. The applicability of the Markov model can thus be evaluated by comparing the attribution levels to the attribution levels of the actual model. The attribution levels are obtained by using the methods described in the subsections in this section.

Although the Markov model is widely used to map customer journeys, it has its limitations. The problem is that we can only use a limited number of states (channels). Increasing the number of channels raises the dimensions of the channel attribution problem, making it infeasible and non-scalable, rather quickly. Furthermore, as mentioned, the idea behind the Markov model is the Markov property. This property states that the probability of transitioning to a future state depends only on the current state and is unaffected by the history of previous states. When applying this to our context, a customer transitioning from one state (channel) to the other, only depends on the channel the customer last visited. Disregarding the user's past interactions may not be a good idea in the highly complex multi-touch attribution problem.

4.2 Random Forest

In this paper a Random Forest model is used to predict whether a given customer's path will end up in a conversion or not. A random forest is a powerful machine learning tool that is part of the ensemble methods, where multiple decision trees are combined to make accurate predictions. The Random Forest model can either be used for regression or classification. We predict whether a path will end up in conversion or not, this means the random forest is used to classify. The thing that makes a random forest model so attractive is the fact that each tree is trained on a random subset of the training data and a random subset of features. This means the algorithm captures diverse patterns and greatly reduces overfitting. During the training process each decision tree is grown recursively. When finally predicting, the Random Forest aggregates the votes from all trees. The final prediction with classification is based on majority voting. In addition to the majority voting, random forests can also provide the probabilistic predictions. The probability of a class label is computed by considering the proportion of decision trees that predict that particular class. The advantage of the random forest model in contrast to the Markov Model is that we are not limited to a number of channels. A random forest is able to handle high-dimensional data.

Now we apply the Random Forest model to the channel attribution problem, using a path as input means more specifically that the frequency of each channel appearing in a path is used as input. For a single input, a random forest is made up of an ensemble of decision trees. A single decision tree in this forest is constructed using a subset of the training data and a subset of the individual channels. To generate the subsets of the training data, bootstrapping is used. Bootstrap sampling involves randomly selecting samples from the original data with replacement. Using random subsets of the channels makes sure correlation is reduced among the trees. During training of the model each decision tree is built using the bootstrap samples and the selected features. A tree gets built recursively by binary splits. At each node, the algorithm evaluates different splitting points for each selected feature. In the case of a random forest for classification, the algorithm chooses the split based on the Gini impurity. The split that results in the greatest impurity reduction is chosen as the best split at that node.

In Figure 2 a general illustration example is shown, it shows that from a certain input X an output Y is obtained. X represents a single path, more specifically it holds the frequencies the individual channels occur in the path. The Figure shows that from X, different trees are built. Each tree is trained on a random subset of the training data and a random subset of features. In every tree the lowest orange node represents the final classification vote. The + sign represents the aggregation of the results, in our case a majority voting for the class label. Finally Y shows the outcome for the single path X, either a 1 for a conversion or 0 for a non converting path X.

4.3 SMOTE sampling

Synthetic Minority Oversampling TEchnique (SMOTE) is an oversampling technique introduced by Chawla, Bowyer, Hall and Kegelmeyer (2002). It is designed to overcome a frequent difficulty

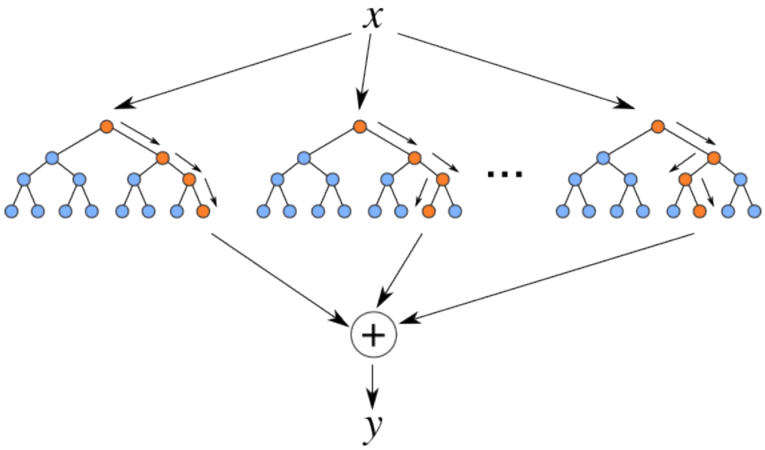


Figure 2: Basic illustration of how a random forest predicts a final value Y, using input X.

in Machine Learning classification algorithms, namely imbalanced data. Imbalanced data refers to datasets where the distribution of classes is significantly imbalanced, with one class being dominant while the other is underrepresented. This problem poses a challenge to many machine learning algorithms, as they tend to bias their predictions towards the majority class. This is a problem as most often we are actually interested in the minority class. SMOTE sampling offers a solution by synthesizing artificial samples for the minority class, thus balancing the dataset.

In this research paper a random forest model is proposed to predict conversion probabilities for individual paths. On a total of 238,679 customer paths only 17,482 paths lead to conversion. This means roughly 7% conversions opposed to roughly 93% paths that do not convert. This means the data is imbalanced, which causes a problem for the performance of the Random Forest model.

The SMOTE algorithm creates new synthetic samples for the minority class in a dataset in the following way: First, the algorithm looks at the instances belonging to the minority class. Secondly, for each minority instance, SMOTE identifies its nearest neighbors in the dataset. Nearest neighbors are instances that are similar to the minority instance in terms of their features. Finally the SMOTE algorithm creates points along line segments that connect the minority instance with its nearest neighbors. These points are used to generate the new synthetic samples.

4.4 Shapley Value

The Shapley Value is a concept that stems from cooperative Game theory, it aims to fairly distribute the total payoff that is generated by a cooperative game among all players based on the individual contributions. It was introduced by economist Shapley (1953) and has shown its use in various fields. Consider a finite set of players (P) each with a fixed strategy. The characteristic function v(X) denotes the value generated by coalition X ⊆ P. The goal is to fairly distribute the total value v(P) among all individual players, based on their contributions. The Shapley Value distributes the value v(P) to a player r ∈ P as follows.

$$\phi_r^{Shap} := \sum_{X \subseteq P \setminus \{r\}} w_{|X|} \times \{v(X \cup r) - v(X)\} \tag{1}$$

$$w_{|X|} := \frac{|X|!(|P| - |X| - 1)!}{|P|!} \tag{2}$$

What makes the Shapley value so attractive, is that it satisfies the following four properties:

1. **Efficiency**: The sum of the Shapley values assigned to all players equals the total value of the game. $\sum_{r \in P} \phi_r = v(P)$

2. **Symmetry**: If two players have the same contributions to any coalition, their Shapley values will also be the same. $\phi_r = \phi_{r'}$ for $\forall r, r' \in P$ s.t. $v(X \cup \{r\}) = v(X \cup \{r'\})$

3. **Linearity**: Consider two characteristic functions $v_1()$ and $v_2()$.
$\forall r \in P$, $\phi_r(v_1 + v_2) = \phi_r(v_1) + \phi_r(v_2)$ and $\phi_r(\alpha v_1) = \alpha \phi_r(v_1)$ for all $\alpha \in \mathbb{R}$

4. **Null player**: If a player does not add any value to any coalition, their Shapley value equals 0. $\phi_r = 0$ if $v(X \cup \{r\}) = v(X)$, $\forall X \subseteq P \setminus \{r\}$

The Shapley Value is particularly interesting as an approach to marketing channel attribution, because it offers valuable insights into the contribution of each channel (player) towards conversion (total value). An individual customer interacts with multiple touchpoints across various channels, before converting. The Shapley value accounts for the incremental value that a single channel brings in influencing the customer behaviour. In contrast to heuristics the Shapley value provides a better understanding of the effectiveness of different marketing channels. This helps marketers identify not only the channels that drive most conversions but also those that have a large impact on the customer's potential conversion. Besides, the Shapley value takes into account interaction effects between channels. This helps to identify combinations of channels that yield the most conversions.

4.4.1 SHAP value

The SHapley Additive exPlanations (SHAP) value is a powerful mathematical method to explain predictions done by machine learning models. The SHAP values can either be used as a local and/or global explainer. Where a local explainer aims to provide explanations for individual predictions or instances, the global explainer aims to provide insights into the overall behavior and functioning of a machine learning model. In this paper, we aim to use the SHAP values as a global explainer, namely explaining overall predictive performance by feature importance. The SHAP value is based on the Shapley value from game theory and provides insights into the importance and impact of different features (channels) that drive the model's predictions. Similar to the Shapley value, the features are considered as players in a cooperative game, where the prediction is seen as the payout. The SHAP values can be used to explain how much a single feature (channel) contributes to the value of a single prediction. In summary SHAP values are a specific implementation of Shapley values, and have specific application in XAI.
The machine learning model in this research is the Random forest model. SHAP values can be applied to the random forest in the following way. Once we have a trained random forest model, the SHAP values for each channel can be calculated for a single customer journey (path) that is used to make a prediction. The SHAP values quantify the contribution each channel has on the model's predicted conversion rate. When finally averaging the SHAP values over

all predicted observations, the overall contribution of each channel in influencing the model's prediction is determined. This value can be interpreted as the overall channel attribution. The exact implemention can be found in Section 4.6.5. So, in conclusion, the SHAP values are interpreted as Shapley attribution levels, as they explain the contribution of a single feature (channel) to a conversion.

4.5 Heuristics

Finally there are some commonly used attribution schemes. They are widely used but have their limitations. In this subsection the Last-Touch Attribution and the Uniform Attribution are explained. The heuristics described in this subsection, are found with the *ChannelAttribution* package (Altomare & Loris, 2023) in R.

4.5.1 Last-Touch Attribution (LTA)

LTA is a marketing attribution model that assigns all the credit for a conversion to the last marketing touch point a customer interacts with before converting. It basically ignores all the preceding touchpoints and credits the final touchpoint as the most influential. This directly shows the limitation of LTA. A path typically consists of several touchpoints, and each of these touchpoints contribute significantly to conversion by generating awereness, interest or consideration. Focusing solely on the last touchpoint leads to skewed understanding of the marketing effectiveness, because the other touchpoints get undervalued or even disregarded.

4.5.2 Uniform Attribution

In the Uniform Attribution model, all credit for a conversion is divided equally over all touchpoints. This approach aims for fairness and does not over emphasize or undervalue a specific touchpoint. It provides a simple model that provides a balanced view of the effectiveness of certain touchpoints. However, there is a limitation of the Uniform Attribution model. By assigning equal weight to every channel, the effectiveness of certain touchpoints is not considered. A more effective channel which requires less attempts to move the customer up to conversion, receives less credit then a channel which needs a lot of attempts to reach the customer towards conversion.

4.6 Analysis

In the Analysis section, the exact methods that were used are explained. First of all, the building processes of the predictive models are explained. Secondly the methods used to obtain the attribution levels are discussed. Finally the algorithms used to obtain the Shapley Values are discussed.

4.6.1 Markov model

The *ChannelAttribution* package (Altomare & Loris, 2023) in R, allows us to create a Markov model. A transition matrix containing the transition probabilities from channel to channel is obtained, the probabilities are estimated using the data, more specifically, on what channels customers interact with. Using the *markovchain* package (Spedicato, 2017) in R, a discrete Markov Chain object is built. This Markov chain allows us to simulate paths. To finally evaluate the performance of the Markov model in predicting the number of conversions, the *ChannelAttribution* package (Altomare & Loris, 2023) is again applied. Via this package the LTA and Uniform attribution are obtained.

4.6.2 Random Forest Model

The random forest model is built in R using the *ranger* package (Wright & Ziegler, 2017). The default settings for creating the forest, namely 500 trees, are used. To train the model we use 70% of our data as training data, which is randomly bootstrapped. The other 30% is used as testing data.

4.6.3 SMOTE sampling

As mentioned in section 4.3, SMOTE sampling is used to get a more balanced dataset. This is necessary because a predictive model has difficulty predicting correctly when working with an unbalanced dataset. The dataset initially contains 17,482 converting paths and 238,679 non converting paths, meaning roughly 7% converting and roughly 93% not converting. Using the *smotefamily* package (Siriseriwan, 2019), synthetic instances of the minority class are generated. To really give the Random Forest model a chance, the conversion rate of 7% of the original data, is changed into a conversion rate of 46.5%, meaning 192,302 converting paths and 221,197 non converting paths. The random forest with SMOTE sampled data, again uses 70% of the SMOTE sample as training data and the other 30% to predict.

4.6.4 Shapley values

To obtain the Shapley values for the individual channels, an algorithm written by Trevor Paulsen[3] is used. The algorithm uses the *GameTheoryAllocation* package (Saavedra-Nieves, 2016).

4.6.5 SHAP values

To obtain the SHAP values for the predicted conversion probabilities, an initial attempt was to use the *DALEX* package (Biecek, 2018). By using the predict_parts() function with *type = "shap"*, the SHAP values for a single prediction are obtained. Because this approach is computationally very expensive for a single observation, another approach is used to get the aggregated SHAP values for the roughly 120,000 other observations. Using the *fastshap* package (Greenwell, 2023), the approximate Shapley values for a set of features are computed. It uses the Monte Carlo algorithm described in (Štrumbelj & Kononenko, 2014). Using 100 simulations, the aggregated SHAP values for the individual features (channels) that contribute to the final prediction (conversion probability) are computed.

[3] See more on: https://trevorwithdata.com/attribution-theory-the-two-best-models-for-algorithmic-marketing-attribution-implemented-in-apache-spark-and-r/.

5 Results

In this section the obtained results are discussed. First of all the results of the actual model is discussed to create a benchmark model to compare the predictive models with. Secondly the results of the simulated data by the Markov model are shown, and the predictions that are done by this model. Furthermore the results of the Random Forest model are discussed, and explained using the SHAP values. Finally the applicability of the predictive models are evaluated by comparing the Shapley attribution levels.

5.1 Benchmark model

From the actual data a benchmark model is initialized. The benchmark model represents the customer paths, with the actual number of conversions per path. The actual number of conversions of a path are obtained by grouping similar paths, and counting the number of paths that ended up in conversion. Table 3 shows the attribution levels per channel. In Figure 3, a barplot is shown. For every channel, the three attribution schemes are shown as a bar. It can be seen clearly, regardless of the attribution scheme, that Social Media contributes the most to a customer conversion. Also, regardless of the attribution scheme, it can be seen that Online Display contributes the least to a customer conversion. Interpreting the Last-Touch results, Social media has a high percentage of occurring as the last step before conversion, this could be interpreted as it being a very effective channel. Comparing to the Uniform, it is seen that the Social Media channel occurs frequently in a customer's path. When looking at the Online Video channel it is seen that relatively it is not frequently the last channel a customer interacts with before conversion. When looking at the Uniform attribution it can be seen that it has a high ratio of occurring in the paths. Finally looking at the Shapley values, which are supposed to give the fair attribution levels. It can be seen that they are almost identical to the Last-Touch attributions. This could be explained by multiple explanations. First of all, it could imply that the paths are generally short or very homogeneous. Secondly, it could mean that attribution in this dataset relies more on effectiveness of certain channels than occurrences.

For replication purposes the original five distinct channels are decreased into four distinct channels. The channels Instagram and Facebook are clustered together, as Social Media. This could also be an explanation for the attribution level of Social Media to be relatively high.

Channel\Attribution scheme	LTA	Uniform	Shapley
Social Media	0.428	0.366	0.427
Online Display	0.121	0.143	0.119
Online Video	0.193	0.275	0.192
Paid Search	0.258	0.216	0.263

Table 3: Attribution levels of the benchmark model.

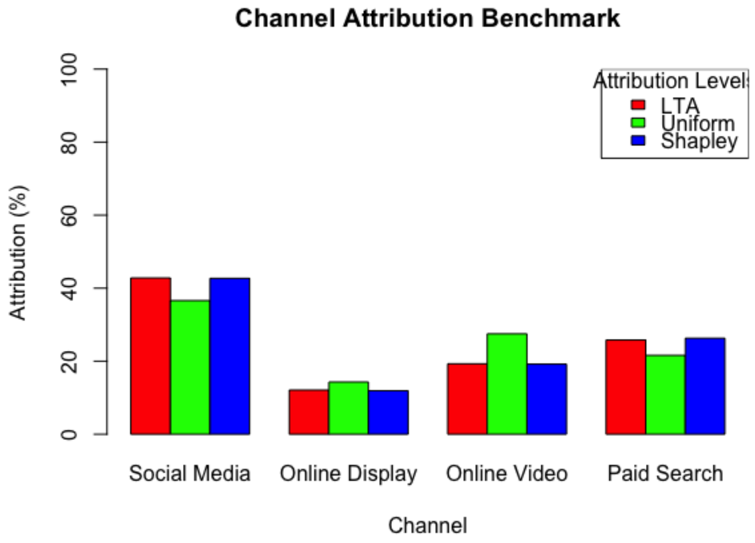
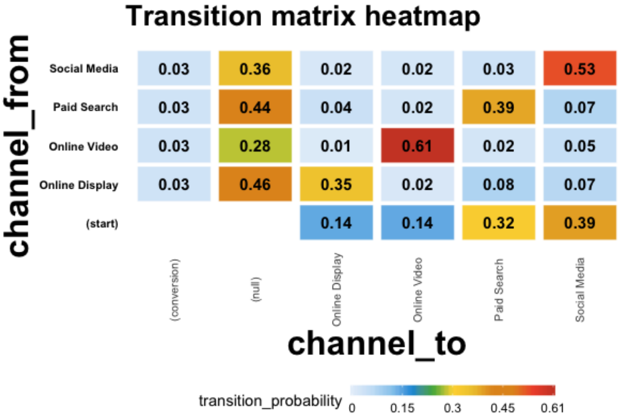


Figure 3: Barplot attribution benchmark model.

5.2 Markov model

To evaluate performance of the predictive Markov model as an underlying model, paths are simulated using the transition probabilities obtained from the actual data. Figure 4 shows a heatmap with the transitioning probabilities in between every state. The states (conversion) and (null), mean the customer leaving the system by buying the product or quitting the system without buying, respectively. Note that a customer has the highest probability of entering the system through social media. This might be a reason for the attribution to be relatively high, compared to the other channels. In Figure 5, a visualization of the Markov Chain is shown. In this Figure, it can clearly be seen that the states (conversion) and (null) are absorbing states. This means that once they are reached, the customer never leaves this state. This is shown by the probability to transition into itself being 1.

In the actual data there are 238,679 individual customer paths, so the exact same number of paths is simulated for the most accurate comparison. Looking at the resulting attribution levels obtained from the simulated Markov data in Table 4, it can be noticed notice that the attribution levels are significantly more equal across the different attribution schemes. The reason for this might be that the focus while simulating paths, lies more on the effectiveness of channels rather than the frequency of channels in a single path. The Markov model relies on the Markov property, which means the history of previous states is disregarded. In this context it would mean that the paths simulated with the Markov property, do not take into account how frequently certain channels have already been visited. In practice this would not be the case. How often a certain channel is visited greatly depends on how many times the channel is already visited. Do note that again Social Media has the highest attribution level, suggesting Social Media contributes the most to the conversion of a customer.



channel_from	(conversion)	(null)	Online Display	Online Video	Paid Search	Social Media
Social Media	0.03	0.36	0.02	0.02	0.03	0.53
Paid Search	0.03	0.44	0.04	0.02	0.39	0.07
Online Video	0.03	0.28	0.01	0.61	0.02	0.05
Online Display	0.03	0.46	0.35	0.02	0.08	0.07
(start)			0.14	0.14	0.32	0.39

Figure 4: Heatmap of transition matrix.

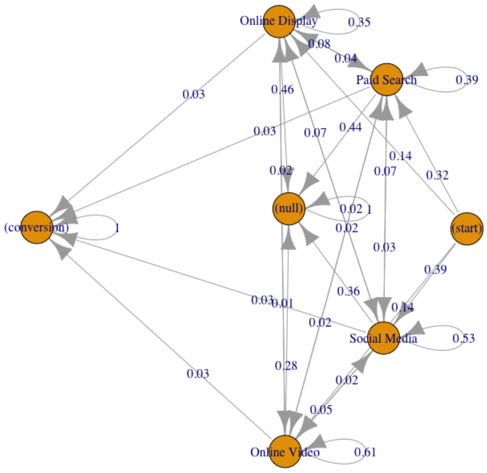


Figure 5: Visualization of the Markov Chain, showing how the customer travels to potential conversion.

Channel\Attribution scheme	LTA	Uniform	Shapley
Social Media	0.425	0.415	0.414
Online Display	0.121	0.122	0.120
Online Video	0.194	0.190	0.199
Paid Search	0.260	0.272	0.267

Table 4: Attribution levels of the model with simulated paths, using the transition probabilities of the Markov model.

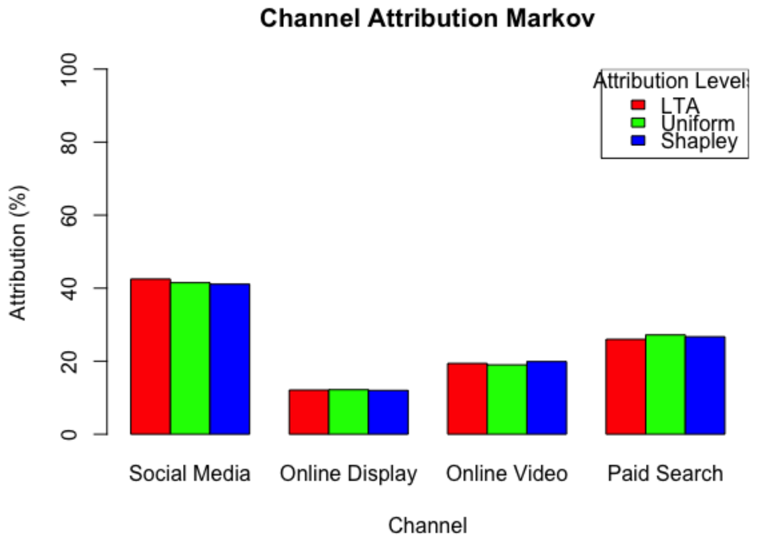


Figure 6: Barplot with attribution levels of the Markov model.

5.3 Random Forest

In this research paper, the aim is to show that a random forest model is a good model to solve the attribution problem. More specifically, the random forest should do a good job in being the underlying model to structure a customer's journey to potential conversion. To show this, the random forest model is used to predict conversion probabilities for a single path. As mentioned in Section 4.6.2, R is used to build a random forest model, using the *ranger* package (Wright & Ziegler, 2017). The model is trained using training data, which consists of a bootstrapped sample of 70% of the entire dataset. Using the random forest model, predictions are made. The predictions are made using the 30% testing data. The output is two columns, the columns represent the probability of the path not converting and the path converting, respectively. The conversion probability is calculated as the proportion of decision trees that votes the path ending up in conversion. To evaluate the performance of the model, and explain the model the SHAP values are used to find the feature importance of the different channels. The SHAP values are used as a global explainer for the predictions made by the random forest. The SHAP values represent the feature importance in the prediction of the conversion probability. The feature importance can be interpreted as the contribution a single channel has to the final value (predicted probability), therefore we interpret the SHAP values as Shapley attribution levels. This also allows us to finally compare the Shapley values of the different underlying models, and thus compare applicability of the underlying models to the channel attribution problem.

In Figure 7, the aggregated SHAP values are shown for every feature (channel). For comparison purposes the feature importance is shown as percentages. The SHAP values for the original random forest model are shown as the red bars in Figure 7. It can immediately be observed that these values are extremely low, which could imply that predictions for the random forest model can simply not be explained. A reason for this is that the predictions of the random forest model are simply not accurate. This could be explained by the fact that the dataset that is used, is highly unbalanced. As mentioned, in order for a predictive model to perform well, it

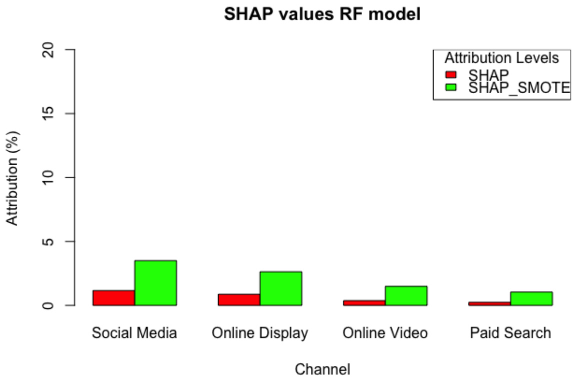


Figure 7: Barplot of SHAP values RF model, with and without SMOTE sampling.

is necessary that the data is well balanced. To give the random forest model a fair chance to show that it is able to predict, SMOTE sampling is used to create a more balanced set.
In Table 5, the ratio of paths that are correctly predicted to convert are shown. It can immediately be seen that for the random forest model using original data this ratio is near zero, implying the predictions are very inaccurate. By using the SMOTE sampling, this ratio has increased to 29%, which implies clear improvement.

	RF model	RF with SMOTE
Percentage correctly predicted conversions	0.17%	29%

Table 5: The percentage of paths correctly predicted to convert, with and without using the SMOTE sample.

With the use of SMOTE sampling, many synthetic paths that end up in conversion, are created. Similar to the original random forest model, the data is split into a 70% training data set, and a 30% testing data set. The trained model again predicts probability of the path converting and of the path not converting. Again we use the SHAP values to globally explain the predicted conversion probabilities. The green bars in Figure 7 show a little increase in SHAP values. A general remark that could be made on the low SHAP values is that they are aggregated for every individual prediction. This implies that the average is taken of the SHAP values over all predictions. In a random forest that classifies, a SHAP value can either be negative or positive. A positive SHAP value implies a positive impact on the model predicting conversion. A negative SHAP value means a negative impact on prediction, thus leading the model to predict not converting. The average of both positive and negative values is close to zero, depending on the exact values.

5.4 Model Comparison

Now all the individual models and their performances have been discussed, the models will be compared. To compare the performance of the individual models, and evaluate their applicability as an underlying model for the channel attribution problem, the Shapley attribution levels of the models are compared. In Figure 8 the Shapley value attributions are shown for every channel

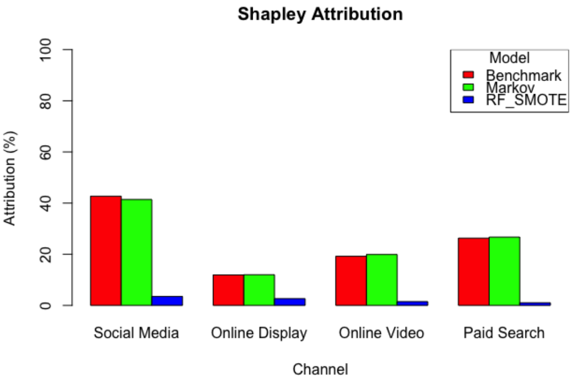


Figure 8: Shapley attribution levels per model.

and model. The Shapley values of the benchmark model are represented by the red bars, the Markov model by the green bars and the Shapley values of the random forest with SMOTE sampling are represented by the blue bars. What can immediately be observed is the fact that the Shapley values of the Random Forest model are relatively low. Where the Shapley values of the other models range from roughly 19% to at most roughly 40%, the Shapley values of the random forest do not exceed 5%. Even when the predictions by the random forest are made using the SMOTE sampled data, the model does not perform well. At least as an underlying model for the channel attribution problem. The random forest might perform alright predictions of the conversion probability, the SHAP values are not a good method to explain Shapley attribution level. A reason for the relatively bad approximated Shapley attribution levels, might be the fact that we used the SHAP values as a global explainer. A global explainer, means that we look at the overall performance and behavior of the predictive model, meaning the SHAP values are averaged over all the predictions. To overcome this problem, a better approach might be to only approximate the SHAP values on the testing data with converting paths. This could allow the SHAP values to only explain the contributions of the features (channels) that pushed the customer towards conversion. This way when averaging the SHAP values over all predictions, there are no negative SHAP values that average the SHAP values to near zero.
Finally, the Benchmark model is compared with the Markov model. It can immediately be observed that the Shapley attribution levels of the two models are relatively close to each other for every channel. This implies that the Markov model performs well as an underlying model for the channel attribution problem, because the Shapley value attribution levels are shown to be very accurate.

6 Conclusion

In this research paper we propose to solve the channel attribution problem using a random forest model. The aim is to demonstrate that the applicability of the random forest model as an underlying model for determining attribution levels of individual channels. This model is intended to overcome the shortcomings of the well known Markov model. The shortcomings of the Markov model are that it is limited to a number of states (channels). As the number of channels increases, the problem will be infeasible. Furthermore, the Markov model disregards important information about the customer's journey, by ignoring the past interactions a customer has had with channels. This neglect results in a loss of valuable information about customer's behaviour, due to the Markov property.

The applicability of the random forest model is evaluated by comparing the Shapley attribution levels of the random forest model with those of the Markov model and the benchmark model. The Shapley values for the random forest model are derived from the SHAP values, which explain the importance of the individual channels on the final predicted probability. These values can thus be interpreted as the Shapley attribution levels.

Our main finding is that when globally explaining the feature importance of the random forest, we obtain limited information about the actual contribution of the features to a final prediction. This limitation arises because the global explanation involves averaging the SHAP values of individual predictions. Since negative SHAP values can occur, the mean Shapley values tend to be close to zero. Consequently, they are not readily interpretable as attribution levels.

In addition to interpreting the feature importance, the initial predictions generated by the random forest model were inaccurate. This was likely due to the use of an imbalanced dataset. However, after employing a SMOTE sampled dataset, the predictions improved and became more reasonable.

All in all, the random forest model is not an appropriate underlying model for solving the channel attribution problem. While it serves as a predictive model capable of estimating conversion probabilities for individual customer paths, the random forest's nature as a black box machine learning model makes it challenging to explain its predictions. The SHAP values were used in an attempt to find the feature importance, and were supposed to be interpreted as Shapley attribution levels. As the SHAP values are considered to perform global explanation on the feature importance they are not well suited as an attribution measure. In the context of the channel attribution problem, the primary objective is to determine the channels (features) that drive customers towards conversion. Therefore, the random forest model is not well-suited for effectively addressing the channel attribution problem.

6.1 Further Research

Based on the conclusion, there are several limitations and potential directions for further research. In this research paper, we solely focus on whether conversion takes place or not. An area of interest is to examine the values conversions generate. Finding what channels drive more valuable conversions, could definitely provide valuable insights.

Additionally, the SHAP values in this paper are used as a global explainer, meaning information

about local predictions are averaged. Local predictions for paths that convert might provide more valuable information, about the importance of individual channels.
Finally, investigating the synergistic effects between different channels in driving conversions, could provide valuable information. Exploring how channels interact in the customer journey, leads to enhanced attribution accuracy and a better understanding of the conversion process.

References

Abhishek, V., Fader, P. & Hosanagar, K. (2012). Media exposure through the funnel: A model of multi-stage attribution. *Available at SSRN 2158421*.

Altomare, D. & Loris, D. (2023). Channelattribution: Markov model for online multi-channel attribution [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ChannelAttribution` (R package version 2.0.7)

Balkanski, E. & Singer, Y. (2015). Mechanisms for fair attribution. In *Proceedings of the sixteenth acm conference on economics and computation* (pp. 529–546).

Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, *19*(84), 1-5. Retrieved from `https://jmlr.org/papers/v19/18-416.html`

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chuang, Y.-N., Wang, G., Yang, F., Liu, Z., Cai, X., Du, M. & Hu, X. (2023). *Efficient xai techniques: A taxonomic survey.*

Greenwell, B. (2023). fastshap: Fast approximate shapley values [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=fastshap` (R package version 0.1.0)

Institute, M. S. (2016, 2017, 2018). *Research priorities 2016–2018.* Cambridge, MA.

Janitza, S., Celik, E. & Boulesteix, A.-L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*, 885–915.

Merrick, L. & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In *Machine learning and knowledge extraction: 4th ifip tc 5, tc 12, wg 8.4, wg 8.9, wg 12.9 international cross-domain conference, cd-make 2020, dublin, ireland, august 25–28, 2020, proceedings 4* (pp. 17–38).

Nisar, T. & Yeung, M. (2015). Purchase conversions and attribution modeling in online advertising: an empirical investigation.

Saavedra-Nieves, A. (2016). Gametheoryallocation: Tools for calculating allocations in game theory [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=GameTheoryAllocation` (R package version 1.0)

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095–1100.

Singal, R., Besbes, O., Desir, A., Goyal, V. & Iyengar, G. (2019). Shapley meets uniform: An axiomatic framework for attribution in online advertising. In *The world wide web conference* (pp. 1713–1723).

Siriseriwan, W. (2019). smotefamily: A collection of oversampling techniques for class imbalance problem based on smote [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=smotefamily` (R package version 1.3.1)

Song, X., Mitnitski, A., Cox, J. & Rockwood, K. (2004). Comparison of machine learning techniques with classical statistical models in predicting health outcomes. In *Medinfo 2004* (pp. 736–740).

Spedicato, G. A. (2017). Discrete time markov chains with r. *The R Journal*, *9*(2), 84–104. Retrieved from `https://journal.r-project.org/archive/2017/RJ-2017-036/index.html`

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*, 647–665.

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi: 10.18637/jss.v077.i01

Yadagiri, M. M., Saini, S. K. & Sinha, R. (2015). A non-parametric approach to the multi-channel attribution problem. In *Web information systems engineering–wise 2015: 16th international conference, miami, fl, usa, november 1-3, 2015, proceedings, part i 16* (pp. 338–352).

Yang, D., Dyer, K. & Wang, S. (2020). *Interpretable deep learning model for online multi-touch attribution.*