Udacity Data Science Capstone Blog Post

Creating a Customer Segmentation Report for Arvato Financial Services

Introduction

In today's hyper-competitive business landscape, understanding your customer base is paramount. In our latest project, we embark on a journey to delve deep into demographics data for customers of a mail-order sales company in Germany. Our mission? To identify the individuals most likely to become loyal customers. To achieve this goal, we draw comparisons between this customer data and the broader population of Germany.

This exciting project unfolds across several stages, each designed to illuminate the path to effective customer acquisition and targeted marketing. Let's take a sneak peek into what you can expect:

Data Exploration and Preparation: Our journey commences with an immersive exploration of the dataset. In this crucial first third of the code, we carefully prepare the data for forthcoming machine learning algorithms. Our tasks include removing columns with missing data, addressing metadata-deficient rows, reformatting columns, eliminating correlated variables, and imputing missing values. To ensure the data is ready for advanced analysis, we also scale the features, priming them for supervised and unsupervised learning to unveil profound insights.

Unsupervised Learning Algorithm: Once our data is meticulously prepared, we dive into the realm of unsupervised learning. Employing Principal Component Analysis, we reduce dimensionality and determine the optimal K-means clusters. This phase sheds light on how the customer population distinguishes itself from the general population, revealing insights cluster by cluster.

Supervised Learning Algorithm: In the final leg of our journey, we delve into supervised learning. Armed with a training set, we prepare the data in a similar fashion as before. Then, we rigorously test various models to gauge their predictive abilities regarding customer likelihood. We leave no stone unturned to identify the top-performing model, fine-tune it, and, ultimately, uncover the best targets for marketing campaigns and enhance customer acquisition. Join us on this exciting expedition to unravel the secrets of effective customer acquisition and targeted marketing campaigns in Germany's diverse landscape.

Part 1: Unlocking the Secrets of Data: Transforming Chaos into Clarity

In this initial stage of our journey, we embark on a mission to unravel the data's mysteries, turning what might seem like madness into a clear path forward. Our goal is simple yet profound: to optimize our ability to predict who is likely to become a customer, ultimately enhancing our business outcomes.

Imagine this stage as our detective work, where we strive to make sense of the wealth of data at our disposal. We want to understand how it can influence the predictability of the models we're about to create—models that will help us identify potential customers. Let's break it down into layman's terms:

Exploring the Attributes: First, we take a peek at the data to understand what attributes exist for both the broader population of Germany and our subset of customers. It's like scanning a big puzzle and examining each piece. We want to know what each attribute represents across the entire dataset.

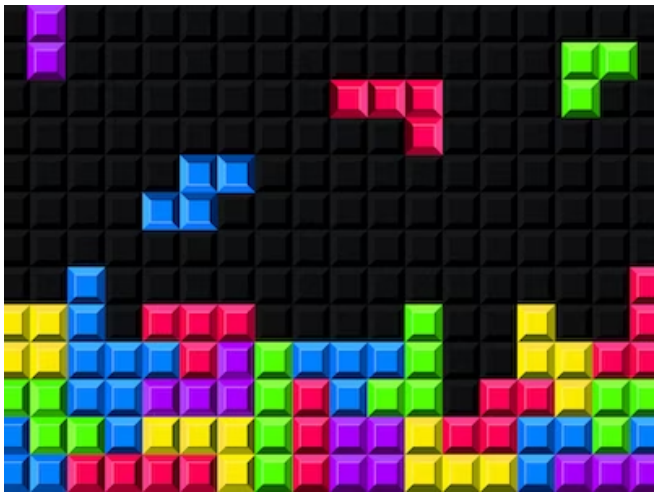
Digging Deeper: We then dig deeper into the details by checking for fields where we have more detailed insights. Think of this as using a magnifying glass to zoom in on certain aspects. This extra layer of information helps us understand the data better. If we lack this additional information, we can't fully comprehend what the data is trying to tell us, so we remove those parts from our analysis.

Being Selective: With a lot of data available, we can be a bit choosy about what we use for our modeling. So, logically, we remove any columns with unclear information. If more than 28% of an attribute's values are missing (like pieces of the puzzle), we remove it to avoid making mistakes during the analysis.

Data Completeness: We also ensure that each row of data is complete. If too many fields are missing (more than 5), we remove the entire record from our dataset. It's like weeding out incomplete puzzle pieces.

Data Enhancement: Once the cleanup is done, we focus on data enhancement to get it ready for modeling. We examine the types of data in each column. If they're not numbers, we either remove them or translate them into a format our model can understand. This might be like converting different currencies into a common unit.

Data Transformation: Some columns needed a makeover. For instance, there were columns representing years, which weren't suitable for direct modeling, so we had to drop them. Other non-numeric fields were converted into numbers, a bit like translating different languages into a common one.



Finishing Touches: Our final steps include creating a correlation matrix to identify and remove any highly correlated variables (imagine avoiding conflicting puzzle pieces). We also fill in missing values through a forward and backfill method, like completing parts of the puzzle that are missing but can be logically guessed. Finally, we scale each attribute to ensure they all fall within a common range, so they play well together in our analysis.

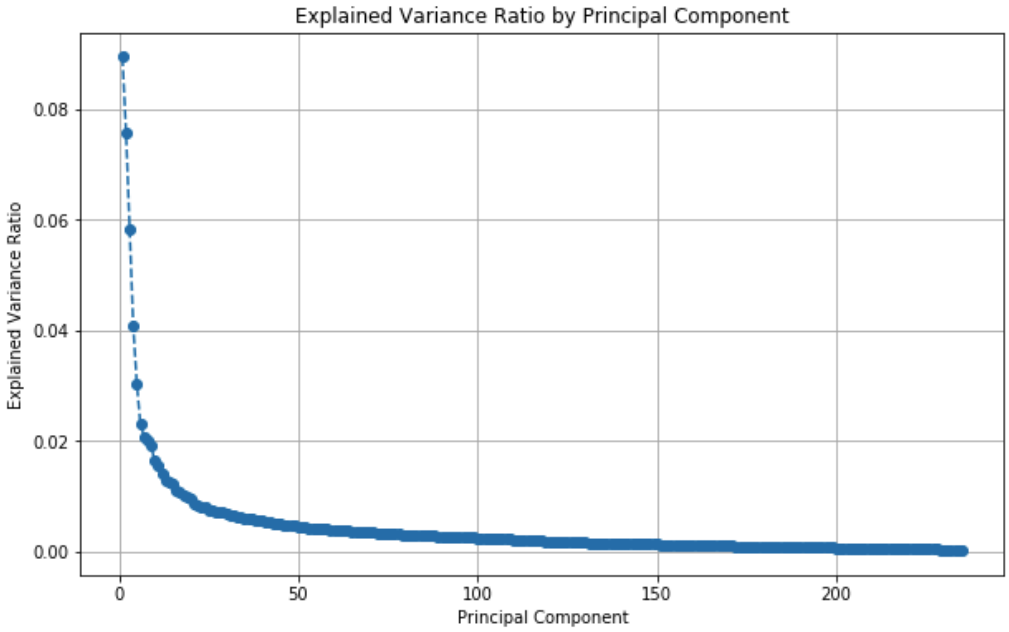
With this stage completed, we've turned the chaos into clarity, and we're now equipped with data that's primed and ready for our upcoming modeling adventures. Stay tuned for more exciting insights in the next phases of our journey.

Part 2: Unveiling Insights with Unsupervised Learning

Now, let's delve into the captivating world of unsupervised learning—a powerful technique that helps us uncover hidden patterns and structures in our vast sea of customer data.

Principal Component Analysis (PCA): Imagine our customer data as a complex web of interrelated attributes, each holding its own piece of the puzzle. Our first step is to simplify this intricate web. We apply Principal Component Analysis, which is like a magic lens that reduces the data's complexity while preserving its essential trends and patterns. It's a bit like turning a tangled jigsaw puzzle into a simpler, more manageable one.

Peering into Variance: We take a closer look at each principal component to see how much of the data's variance it explains. Think of it as understanding how significant each piece of the puzzle is in telling the whole story. We strive to determine the optimal number of components needed to capture the most meaningful variance.

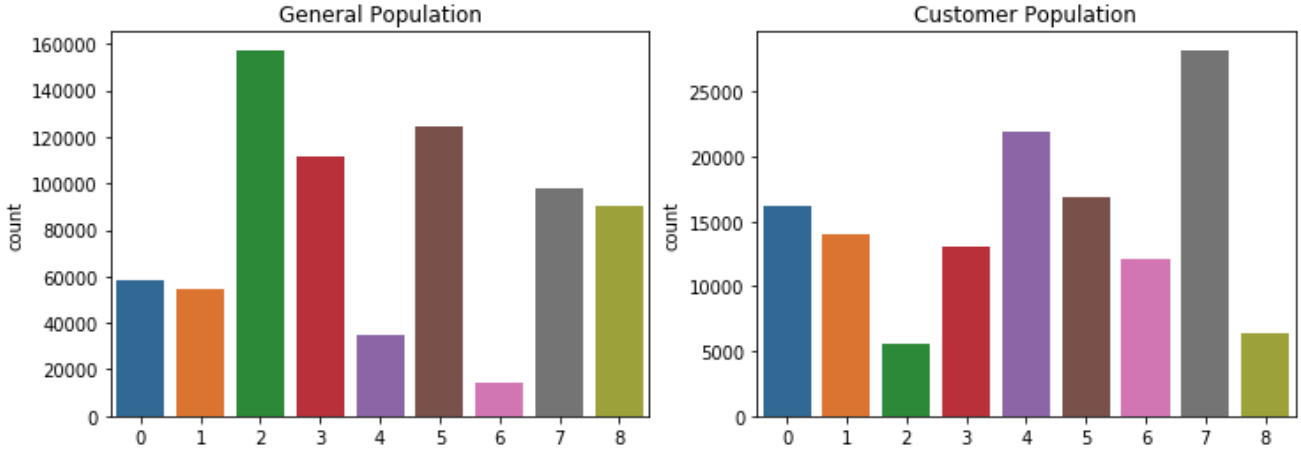


Aim for Clarity: Armed with this number, we reapply PCA to our data. Our goal? To ensure that our analysis retains at least 85% of the data's variance. It's like ensuring that the simplified puzzle still contains enough pieces to paint a clear picture. If it does, we're ready to move on to the next exciting step.

K-Means Clustering: Picture our data as a vast universe of customers, and within this universe, we aim to discover distinct communities or clusters. K-Means Clustering is our spacecraft for this mission. We test various cluster sizes, from as few as 1 to as many as 21, observing how our data divides into groups. This technique helps us identify patterns, compress the data, pinpoint outliers, and make better sense of the customer landscape.

Finding the Perfect Cluster Count: We're on a quest to find the Goldilocks cluster size—not too few, not too many, but just right. To do this, we minimize the square sum of errors and visually inspect how the cluster count affects our data. It's like choosing the right magnification for viewing intricate details.

Comparing Universes: Once we've determined the ideal cluster count, we take a side-by-side look at how clusters differ between the general population and our customers. This reveals where we might over or underrepresent certain groups, providing valuable insights for our marketing and customer acquisition strategies.



Unsupervised learning is our secret weapon for making sense of the data's complexity, finding the right clusters, and unlocking the hidden patterns within our customer universe. Stay tuned for the next phase of our journey, where we delve into supervised learning to refine our predictive abilities.

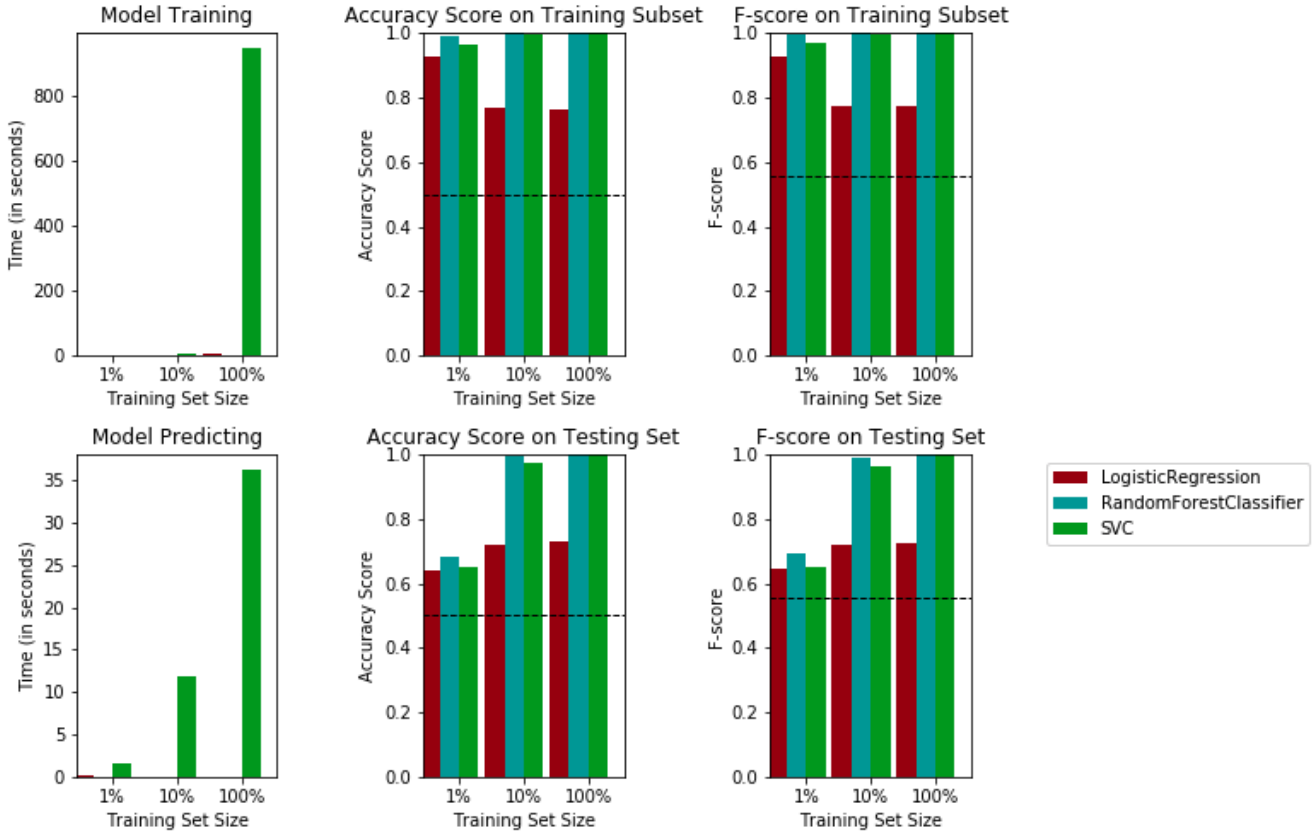
Part 3: Supercharge Your Predictions with Supervised Learning

In the thrilling phase of our journey, we introduce the power of supervised learning to train our model. This model will guide us in understanding customer behavior, helping us tailor our strategies for maximum impact.

Understanding the Training Data: As we step into the world of supervised learning, our first task is to gain insights from a fresh training dataset. This data informs us on how to clean and optimize the information we'll feed into our predictive model. We start by examining how customers have responded in this training set, primarily focusing on the column that distinguishes those who became customers from those who didn't. In some cases, these numbers are uneven, with just a tiny fraction—about 1.5% of the population—becoming customers. To level the playing field, we employ a resampling technique to ensure a more balanced distribution of responses. It's like giving each piece of the puzzle equal weight.

Data Cleanup and Preprocessing: With resampling complete, we apply all the data cleanup and preprocessing steps we've detailed previously. We whip the data into shape, ensuring it's ready to work its magic in the modeling phase.

Selecting Models: The heart of this phase involves choosing the right models to work their predictive wonders. We test three models—Logistic Regression, Random Forest Classifier, and Support Vector Classifier (SVC)—to understand how they perform with varying sample sizes.



Optimizing the Model: Once the models have had their chance to shine, we compare their results. In this scenario, both Random Forest Classifier and SVC prove highly effective. Random Forest Classifier takes the lead, and we opt to go with it for further training. Though further training isn't strictly necessary (given that RFC achieved a perfect prediction rate), it's still wise to fine-tune and optimize our parameters. This ensures we're running at peak efficiency and can handle any uncertainties that may arise in a real-world testing set.

Peering into the Magic Box: As we wrap up this phase, we take a quick peek at the normalized weights for the five most predictive features. This is like opening the magic box to see which variables have the most influence in making predictions.

Testing the Model: In the final stretch, we introduce a brand new testing dataset to our model. We clean it up, feed it into our finely tuned model, and watch as it works its predictive wonders. From here, we can precisely target individuals most likely to become customers, turning them into loyal patrons of our business.

Conclusion

In the grand culmination of our data-driven odyssey, we've journeyed from unraveling the mysteries within the raw data to supercharging our predictive prowess. With unsupervised learning, we've uncovered hidden patterns and structures, turning data chaos into clarity. Then, as we entered the world of supervised learning, we honed our predictive model to perfection, optimizing its accuracy and efficiency. Now, armed with this formidable tool, we're poised to revolutionize our approach to customer acquisition and targeted marketing. The knowledge we've gained from each stage of our journey will serve as the compass guiding our strategic decisions and steering us toward a future filled with successful endeavors.