Introduction to ML for Business

Case Study Worksheet: Startup Profiling from Investments

Group:
Date:
Course: Machine Learning for Business

Problem Representation

How can you profile a startup using investment data?

Identify and streamline the types of data that are critical for a thorough analysis to understand the factors that influence startup success and investment attractiveness. Consider the impact of different data types and how they can reveal insights into startup viability, market trends and investor behavior.

Consider a situation where investors are evaluating startups for potential investment. How should they approach this data-driven analysis? Consider variables such as industry, funding history, and growth metrics in your evaluation.

Data Source

Dataset Description: The StartUp Investments (Crunchbase) dataset is a comprehensive collection of data on startups and their investments, sourced from Crunchbase.

Dataset Features: It includes detailed information such as startup names, sectors, funding rounds, investment amounts, and investor details. This dataset is useful for analyzing investment patterns, identifying characteristics of successful startups, and understanding investor behavior.

Access the dataset: To access this dataset, visit the Kaggle link provided.

Here is an example of three dataset records.

name	homepage_url	funding_total_usd	status	country_code
.Club Domains	http://nic.club/	70,00,000	nan	USA

name	homepage_url	funding_total_usd	status	country_code
.Fox Networks	http://www.dotfox.com	49,12,393	closed	ARG
0-6.com	http://www.0-6.com	20,00,000	operating	nan

Task Identification

In the context of analyzing customer behavior for an online retail environment, which research approaches would provide the most insightful understanding of customer buying trends and behaviors?

What are the advantages of using machine learning techniques over traditional statistical data analysis methods?

Data Exploration

Key Variables: Identify key variables in the data set that may have a significant impact on customer retention.

Data Limitations: Examine the data set for any limitations that could potentially affect our analysis. Specifically:

- Are there any instances of *missing values* within the data set?
- Are the *missing values* associated with key variables that are critical to our analysis?
- Do the values within the dataset *exhibit deviations from typical behavior* that could potentially affect our analysis?

Data Preparation for Machine Learning

In the context of preparing the dataset for a machine learning model, please discuss the following key issues:

- How would you handle missing data points?
- Describe your strategy for *feature engineering*. Would you create derived metrics or new variables to improve the predictive performance of the model? If so, provide examples of these derived metrics and their relevance to the analysis.

Model selection

In the context of predicting customer behavior using machine learning, please provide insights on the following:

Classification model: If you are treating the problem as a *classification task*, describe:

- The features included in the training data set.
- The algorithm used to train the model.
- The target used for classification.

Regression model: If you are considering a *regression approach*, describe

- The features included in the training data set.
- The algorithm used to train the model.
- The target to be predicted by the regression model.

Clustering model: In the case of a *clustering model*, explain:

- The features used to cluster customers.
- The algorithm used to train the model.
- The nature of the clustering result. What does each cluster represent? How many clusters do you expect?

Model Evaluation

Training and Testing the Model: Describe the process for training and testing the model. How will you allocate data (i.e. split) for *training and testing* ? Are there any special techniques you plan to use for *hyperparameter tuning*?

Interpreting Results: Explain your approach to *interpreting the performance* of the model. What *metrics* will you use to evaluate the model's predictive ability (e.g. accuracy, precision, recall) and *why*?

Business Insights

Customer Segmentation Based on Spending Behavior Metrics: The instructor conducted an analysis of customer spending behavior in the dataset. This analysis involved deriving specific metrics from the data and using these metrics to classify customers into different segments. An example of the data used for this analysis is shown below.

The metrics used for segmentation are as follows:

- Total Spend: The total amount spent by each customer.
- Frequency: The frequency of the customer's purchases.

- AvgSpendPerInvoice: The average spend per invoice.
- SpendCategory: A categorical label indicating whether a customer is a “high spender” or a “low spender.
- UniqueStocks: The number of unique items purchased by each customer.

Customer ID	TotalSpend	Frequency	AvgSpendPerInvoice	SpendingCategory	UniqueStocks
12346	-64.68	17	-3.80471	Low Spender	30
12347	5633.32	8	704.165	High Spender	126
12348	2019.4	5	403.88	High Spender	25
12349	4404.54	5	880.908	High Spender	139
12350	334.4	1	334.4	Low Spender	17

Rules for Classifying Customers: The results of the analysis led to the formulation of rules for classifying customers into spending categories by using **a decision tree algorithm**. These rules were visually presented in the “Rules for Classifying Customers” figure, which provided a clear guideline for categorizing customers based on their spending behavior.

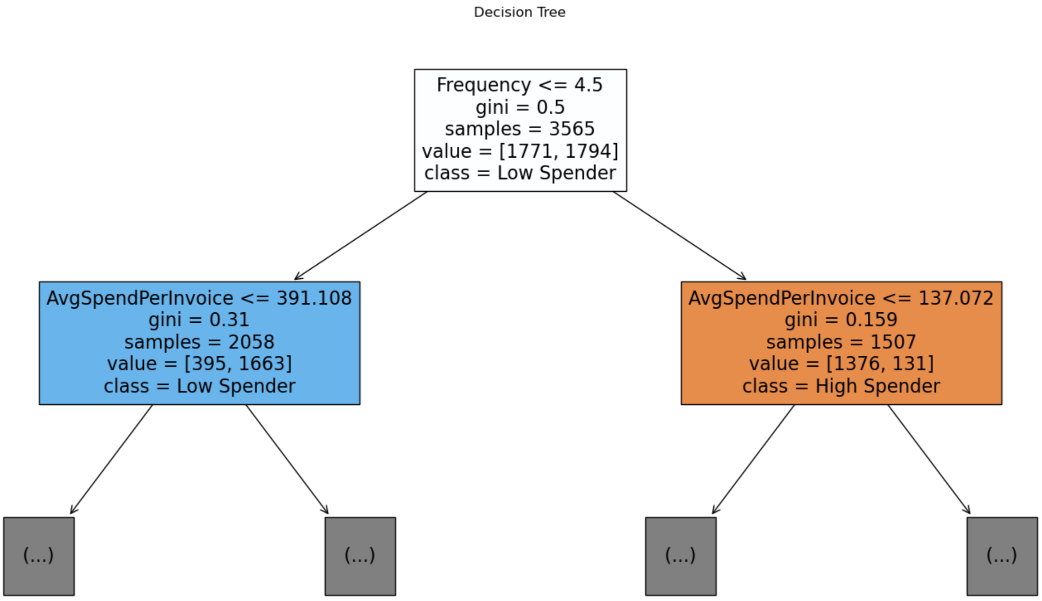

Figure 1: Rules for Classifying Customers

Feature Importance Ranking: To determine whether a customer fell into the “high spender” or “low spender” category, the instructor used a feature importance ranking. This ranking was visually represented in the Feature Importance chart, which showed the importance of each metric in making this determination.

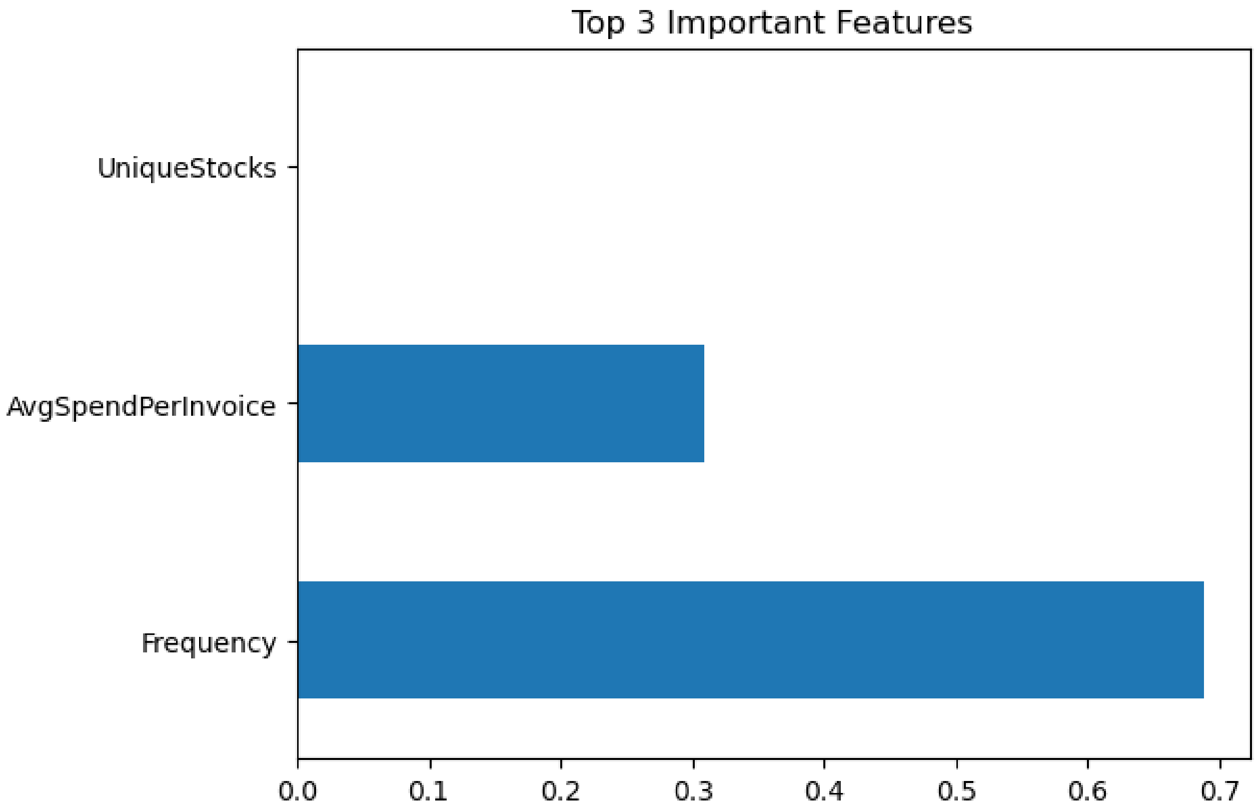

Figure 2: Feature Importance

Derive business insights: *What valuable business insights can be derived from the predictions generated by the model?*

Recommendations for Strategic Actions: *Based on these insights, what specific strategic actions or initiatives can the store take to increase customer loyalty and drive business growth?*

References

- UCI Machine Learning Repository
- IBISWorld
- Journal of Retailing
- McKinsey & Company