Comparing Feature Selection Methods: Traditional (PCA, tSNE, LDA) VS Causality-based

Team #8
MSc Katarina Milenković (katarina.milenkovic@student.tugraz.at)
MSc Matej Vuković (matej.vukovic@student.tugraz.at)

Motivation
When trying to solve a machine learning problem and train a machine learning model, one often deals with highly dimensional sets of data with a lot of features. The challenge is to distinguish between relevant and irrelevant features and to select the minimal subset of relevant features which have an impact on the target label. Choosing the proper subset is a crucial when training a model. Poorly chosen feature subset or subset with features that have no impact on a target label can significantly impair the performance of a model and decrease its accuracy. This process is called feature selection. Nowadays researchers argue1 that feature selection often represents the primary goal of data analysis.

Traditional feature selection methods select features based on the statistical analysis (mathematical, probabilistic, etc), while causality-based methods look at the causal links between the features and the target variable. The goal of our work is to compare and evaluate the two approaches.

Dataset Description
As a working example for the project, we choose Pump it Up: Data Mining the Water Table dataset2. It’s a dataset from Taarifa and the Tanzanian Ministry of Water. The dataset contains different variables that describe a water pump: what kind of pump is it, when is it installed, who installed it, where it is installed, etc. The target label describes the status of the pump, is it functional, does it need some repairs, or is it not functional. The possibility to predict all the failures and maintenance needs on time could significantly improve the water supplying in Tanzania.

Cleaning and preprocessing data
Data cleaning implies the correction of different data quality problems. It’s necessary to have a good understanding of a problem that you are dealing with, in order to do perform this step properly.

The methods we used to overcome the missing values problem are:
- ignoring missing observations/features,
- imputing missing data using some neutral value,
- imputing missing data using cold-deck (mean/other statistic analysis value),
- imputing missing data using hot-deck (the most similar observation),
- imputing missing data using prediction.

The other big problem with our dataset was high cardinality. To overcome this issue we aggregated feature categories into smaller number of categories - leave the minimum number of unique categories needed to account for roughly 80% of all observed categories3.

Encoding the categorical values was done using:
- Target encoder
- Hashing encoder

To prepare the data for the classification algorithms sensitive to outliers or ones relying on distances, standardization/normalization was done using several approaches implemented in scikit-learn library: RobustScaler, PowerTransformer, etc.

Traditional feature selection
First, we established a baseline accuracy by testing the performance of 7 different algorithms on 16 differently preprocessed datasets. Two best baseline models, RandomForest and AdaBoost, were taken for further optimization using traditional feature selection methods:

- Distributed Stochastic Neighbor Embedding (t-SNE)
- Principal Component Analysis (PCA) - mathematical technique
- Linear Discriminant Analysis (LDA) - probabilistic technique

We tested the effectiveness of the above methods on two datasets: hash encoded dataset scaled by PowerTransformer with 18 features (Graph 1) and target encoded dataset scaled by RobustScaler with 33 features (Graph 2).

Causality-based feature selection
As shown in Kui et al., there have been significant developments in the domain of causality-based feature selection. We tried several approaches with mixed outcomes due to the limited computing power and the nature of our dataset. Finally, best results were achieved using following algorithms:
- MMBB (Max-min Markov blanket)
- MMPC (Maximum Minimum Parents and Children)8

Algorithms work in a similar way with MMBB being more complex as it returns the complete Markov blanket of a target value.

Features that were identified as most important: installer, construction_year, wp_name, ward, extraction_type, management, payment, quantity, and waterpoint_type. Best classification accuracy score with these features was achieved with RandomForest ensemble method on a Target Encoded Dataset scaled using RobustScaler. Accuracy of the model is: 0.7970.

Evaluation
Table below shows the comparison of best results (accuracy) from both traditional and causality-based feature selection approaches. Approaches achieved comparable results. More specifically, traditional methods have shown the third decimal advantage.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Original</th>
<th>Other values (2966)</th>
<th>(Missing) 2316</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>5101</td>
<td>False</td>
<td>5055</td>
</tr>
<tr>
<td>False</td>
<td>3334</td>
<td>Missing</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Unique before</th>
<th>Unique after</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding</td>
<td>1898</td>
<td>96</td>
</tr>
<tr>
<td>Installer</td>
<td>2146</td>
<td>102</td>
</tr>
<tr>
<td>WP Name</td>
<td>37400</td>
<td>25517</td>
</tr>
<tr>
<td>GPG</td>
<td>125</td>
<td>71</td>
</tr>
<tr>
<td>Ward</td>
<td>2092</td>
<td>957</td>
</tr>
</tbody>
</table>

I-SNE method is primarily used for visualization and it achieves good accuracy for 2 and 3 features. PCA achieves the best classification accuracy score of 0.7830 (0.8038 after hyperparameter tuning) with RandomForest classifier on the subset of 29 features. LDA achieves the best classification accuracy score of 0.7607 (0.7686 after hyperparameter tuning) with AdaBoost classifier on the subset of 2 features.

However, looking only at the accuracy score is not enough. Changes in the models and the amount data used for training these models are significant. PCA showed competitive accuracy score with around 29 components in contrast with MMPC approach that used 9 features. LDA approach produced a constant accuracy score that was significantly lower than one from other approaches.

References
2. https://livebook.datascienceheroes.com/data
3. https://www.drivendata.org/competitions/7/pump
4. https://www.tugraz.at/
5. https://www.tugraz.at/