DataRobot Feature Discovery

Overview
"Some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used"

Pedro Domingos, The Master Algorithm
# Data Schema

## USER
- **User_id**
- **Time**
- will buy (bananas)

## ORDERS
- **User_id**
- **Order_id**
- **Order Time**

## TRANSACTION
- **Order Time**
- **Reorder**
- **Order_id**
- **Product_id**
- **Product_name**
- **Aisle**
- **Department**

### Link to full open source instacart dataset
Feature Engineering is Laborious and Time-Consuming

What features can you think of to create for buying a product, say bananas?

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>70</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>70</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>70</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>82</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>82</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>82</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>82</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>82</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>109</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>109</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>109</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How many items they usually order? Did they buy oranges? Do they shop in the fruit aisle?
Building a training dataset usually requires to...

1. Define the unit of analysis and the target (in a primary dataset).

2. Perform feature engineering from secondary datasets to enrich the primary dataset.
Feature engineering is hard

Until now, customers had to do feature engineering outside of DataRobot and we usually get the follow feedback

Complex
“I don’t know what transformations are the right ones for my use case”
“How should I aggregate these transactions? Using Sum? Avg? Avg in last week? ..."

Time consuming
“Experimenting many transformation takes so much time that I end up not using my transactional datasets.”
“IT can’t use my Python script in production, I have to replicate the feature engineering code in SQL”

Error prone
“I sometimes forget to make sure to only use feature values that were known prior to each event”
“Our feature engineering process is hard to audit when we get wrong results”
DataRobot Feature Discovery solves this problem

Users can now automatically discover features from secondary datasets to get additional predictive value from features that will be used to train models and make predictions.
Feature Discovery  |  Find Predictive Signal In Related Data.

Automatically discover valuable features in primary and secondary datasets.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Index</th>
<th>Importance</th>
<th>Var Type</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>order_products[reo...red] (30 days sum)</td>
<td>59</td>
<td>59</td>
<td>Numeric</td>
<td>153</td>
</tr>
<tr>
<td>order_products[ais...ays most frequent]</td>
<td>20</td>
<td></td>
<td>Categorical</td>
<td>84</td>
</tr>
<tr>
<td>order_products[rear... ] (30 days counts)</td>
<td>52</td>
<td></td>
<td>Summarized Categorical</td>
<td>1,259</td>
</tr>
<tr>
<td>order_products[pro...at] (30 days sum)</td>
<td>41</td>
<td></td>
<td>Numeric</td>
<td>449</td>
</tr>
<tr>
<td>order_products (30 days num)</td>
<td>7</td>
<td></td>
<td>Numeric</td>
<td>146</td>
</tr>
<tr>
<td>order_products[reordered] (1 week avg)</td>
<td>42</td>
<td></td>
<td>Numeric</td>
<td>257</td>
</tr>
<tr>
<td>order_products[reo...ek most frequent]</td>
<td>47</td>
<td></td>
<td>Numeric</td>
<td>2</td>
</tr>
<tr>
<td>orders[order_time] ... (30 days counts)</td>
<td>94</td>
<td>94</td>
<td>Summarized Categorical</td>
<td>167</td>
</tr>
<tr>
<td>time (days from or...me) (30 days std)</td>
<td>112</td>
<td></td>
<td>Numeric</td>
<td>1,579</td>
</tr>
<tr>
<td>orders[order_time] ... (30 days entropy)</td>
<td>81</td>
<td></td>
<td>Numeric</td>
<td>31</td>
</tr>
<tr>
<td>time (binned hours ... (30 days entropy)</td>
<td>101</td>
<td></td>
<td>Numeric</td>
<td>15</td>
</tr>
</tbody>
</table>

DataRobot allows you to mix location features with numerical, categorical, text and image features in the same dataset.

DataRobot new feature type represents aggregated categorical data.
Feature Discovery | Find Predictive Signal In Related Data.

Trace the feature lineage for any automatically generated feature.

Next-Level Automated Feature Engineering
Feature Discovery | Find Predictive Signal In Related Data.

Friendly workflow to declare relationships between multiple datasets prior to modeling.

Relationship editor makes it easy to declare joins, set prediction points, and see feature derivation windows.

Next-Level Automated Feature Engineering
Feature Discovery | Find Predictive Signal In Related Data.

Ability to download feature derivation log and your auto-enriched dataset.

Integration with model deployments for reliable predictions

Next-Level Automated Feature Engineering
Questions & Answers
Engage, learn, and accelerate your AI/ML journey

Connect with peers to find solutions to AI challenges

Explore helpful content to take your AI to the next level

Build your brand as an AI expert & thought leader

Join your peers today at community.datarobot.com

Questions: learning_sessions@datarobot.com
DATA ROBOT ENTERPRISE AI PLATFORM | RELEASE 6.2

ONE PLATFORM FOR ALL USERS

CREATORS
- Data Engineers
- Data Scientists
- Business Analysts
- Software Developers

OPERATORS
- DevOps
- IT
- Risk and Compliance

CONSUMERS
- Business users
- Senior Executives

CREATE
Prepare, Build, Test

OPERATE
Deploy, Monitor, Manage

CONSUME
Integrate, Use, Benefit

END-TO-END AUTOMATION

GOVERNED, TRUSTED AND EXPLAINABLE AI

MULTIMODAL

DEPLOY WHERE YOU CHOOSE

On-Prem, VPC, Hybrid or Fully Managed Cloud
Multi-Cloud

Security
Collaboration
Explainability
Lineage

DataRobot