P07: Modelling fast growing companies

Team members: Roland Pajuleht, Lisanna Lehes, Simo Jaanus, Kertu Nurmberg
Lennart Kitt
- Head of Customer analytics and Data Science / Data Office @ SEB Baltics

Kristina Lillo
- Innovation Lead
PROBLEM

- Analyze available open data from the Tax Board and Statistics Estonia
- Find a way to identify fast-growers
- Select attributes and build a model

Many new companies are being established monthly.

Bank needs a faster way to find fast-growing companies to provide daily banking solutions and financing.
DATA LABELLING

1. Average Increase

\[ \text{averageIncrease} = \text{avg}(\frac{\text{Revenue}_{i+1} - \text{Revenue}_i}{\text{Revenue}_i} \times 100\%) \]

2. Compound Quarterly Growth Rate

\[ \text{CMGR} = \frac{\frac{1}{\text{Last quarter} \#\text{ of quarters difference}}}{\text{First quarter}} - 1 \]
Estonian Tax Office. Quarterly turnover and tax payment info on company level

Statistics Estonia. General information about companies

List of tax debtors.

Nr of rows: 631 635
Unique companies: 60 326

Merged datasets.

Created $n$ rows of data from each company where $n$ is the number of quarters the company has reported data. Each subsequent row of data has 1 less quarter of data than the previous.

Set the initial cut-off point for \textit{new} companies $\rightarrow$ dropped companies which were created before 2015.

Added \textit{new variables}. Such as whether a company is bankrupt or being liquidated.

Assigned \textit{labels} using formulas (CMGR was chosen).

Created train, test, validation data.

Researched about what has been done in the same field before $\rightarrow$ chose models accordingly.

Tried different \textit{models}:
- Logistic Regression
- ExtraTrees Classifier
- RandomForestClassifier
- Bagged DT
- XGBoost
- Deep learning

\texttt{GridSearchCV}

\textbf{Performance} measured with:
- accuracy,
- AUC,
- F1-score.
<table>
<thead>
<tr>
<th>Kood</th>
<th>Maksevälja</th>
<th>Vaidlustat</th>
<th>Ksuvõlgad</th>
<th>Tähtkood</th>
<th>8-kohajaestli jaoks kall Eestl. Päästeeemäärat</th>
<th>kustutat või lisamüü</th>
<th>pankrotits</th>
<th>korrdeeltarajandmise</th>
<th>label</th>
<th>label</th>
</tr>
</thead>
</table>
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 129.7102 | 17.7574 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 129.7102 | 17.7574 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 127.0556 | 12.7598 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 136.8556 | 14.6876 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 152.9601 | 23.7844 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 48.28278 | 11.9507 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 19.42649 | -0.10579 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 26.12576 | 6.635387 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 54.41807 | 15.75741 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 27.30348 | 9.66498 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 15.20002 | 1.15344 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 19.86919 | 3.988084 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 30.37662 | 13.15741 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 46.62275 | 30.01306 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 42.68542 | 24.31806 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 42.68542 | 24.31806 fast-growing  
| 3308.35 | 2556.72 | 10930.91 | 2 | 6714.56 | 3488.56 | 24875.35 | F | 4329 S11002-S1 | 8 ********** | 0 | 42.68542 | 24.31806 fast-growing  

- F: fast-growing  
- not: not-fast-growing
RESULTS

- Notebook running time ~ 1h
- Bagged DT - acc 95.9%
- XGBoost - acc 96.1%
RESULTS

AUC scores

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>0.98</td>
</tr>
<tr>
<td>BaggingClassifier</td>
<td>0.955</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>0.877</td>
</tr>
<tr>
<td>ExtraTreesClassifier</td>
<td>0.634</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.595</td>
</tr>
<tr>
<td>Deep learn</td>
<td>0.499</td>
</tr>
</tbody>
</table>

F1 scores

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>0.956</td>
</tr>
<tr>
<td>BaggingClassifier</td>
<td>0.955</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>0.881</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.879</td>
</tr>
<tr>
<td>ExtraTreesClassifier</td>
<td>0.878</td>
</tr>
<tr>
<td>Deep learn</td>
<td>0.878</td>
</tr>
</tbody>
</table>
LESSONS LEARNED

- Applying theoretical knowledge learned during course on a real-life problem
- Data preprocessing really is the most important part
- Although working with only open data, it still enabled to find satisfactory models
- Colab might not be the most reliable tool to use
Thank you for listening!

Questions?

Project in GitHub