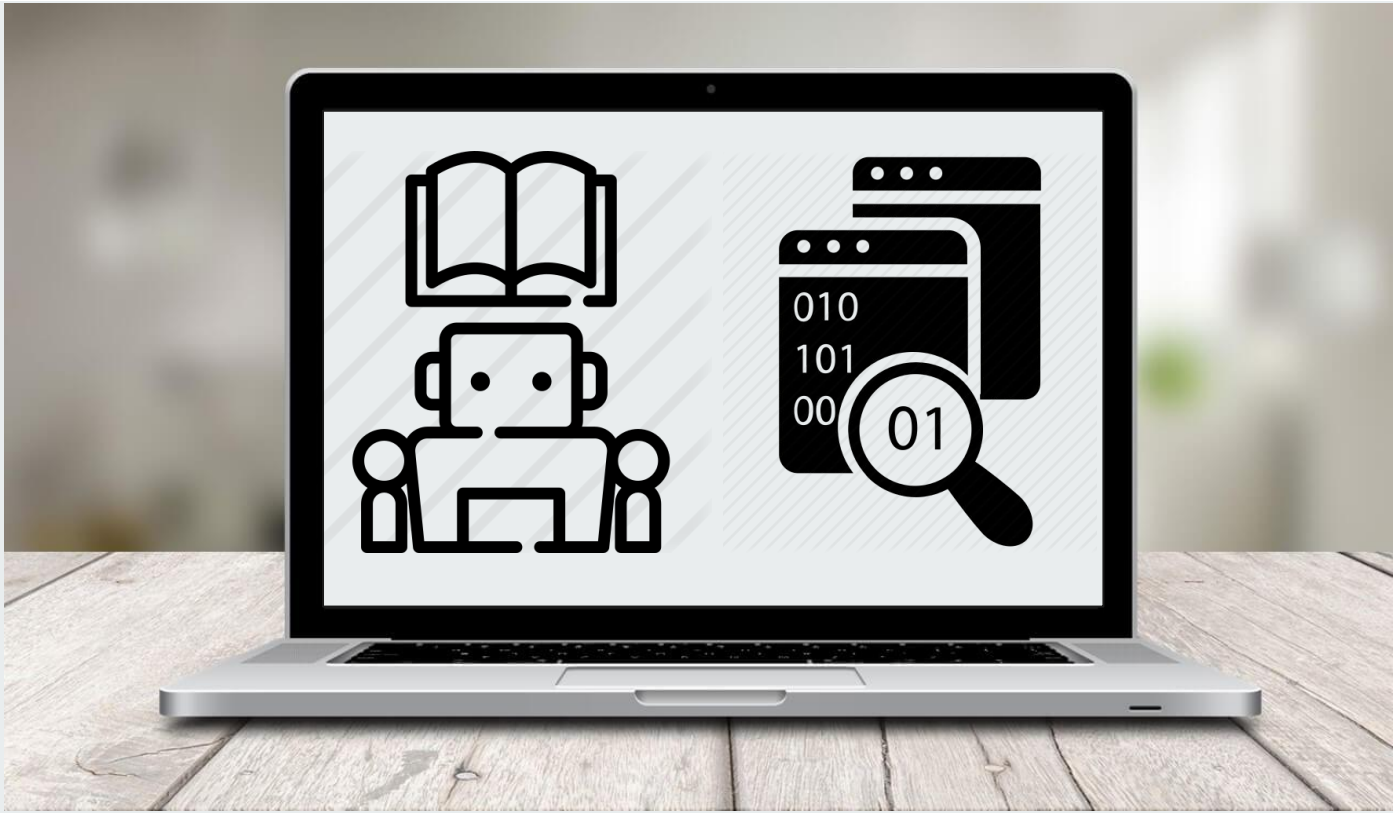
Alberto Marengo

Text to SQL



Using ML to generate queries from Natural Language

The Question

Can AI read a Natural Language Query and translate it into a Database Query?

Data Description

wikiSQL ~56,000 questions and SQL queries

Question:
Who played in the Toronto Raptors from 1995-1996?

SQL query:

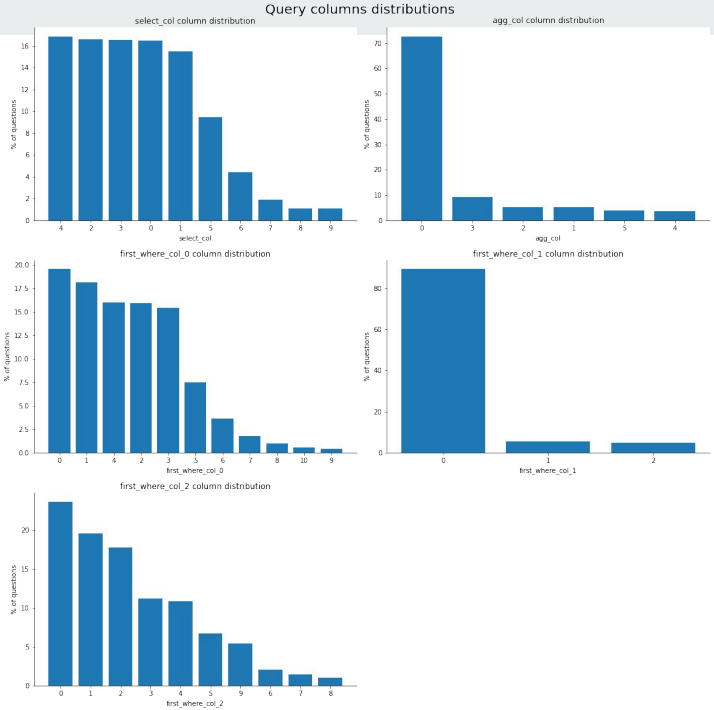
SELECT column_0
FROM 1-10015132-14
WHERE column_4 = “1995-96”;

SELECT column #

WHERE statement
1. Column #
2. Operator → 0→ = 1→ > 2→ <
3. Match

GROUP BY operator
0→ no aggregation
1→ MAX
2→ MIN
3→ COUNT
4→ SUM
5→ AVG

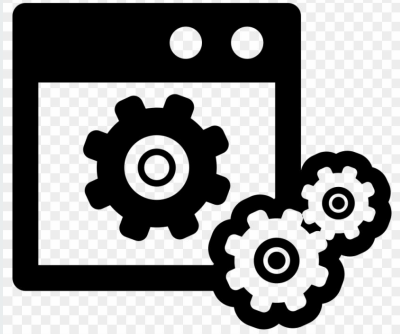
Data Description



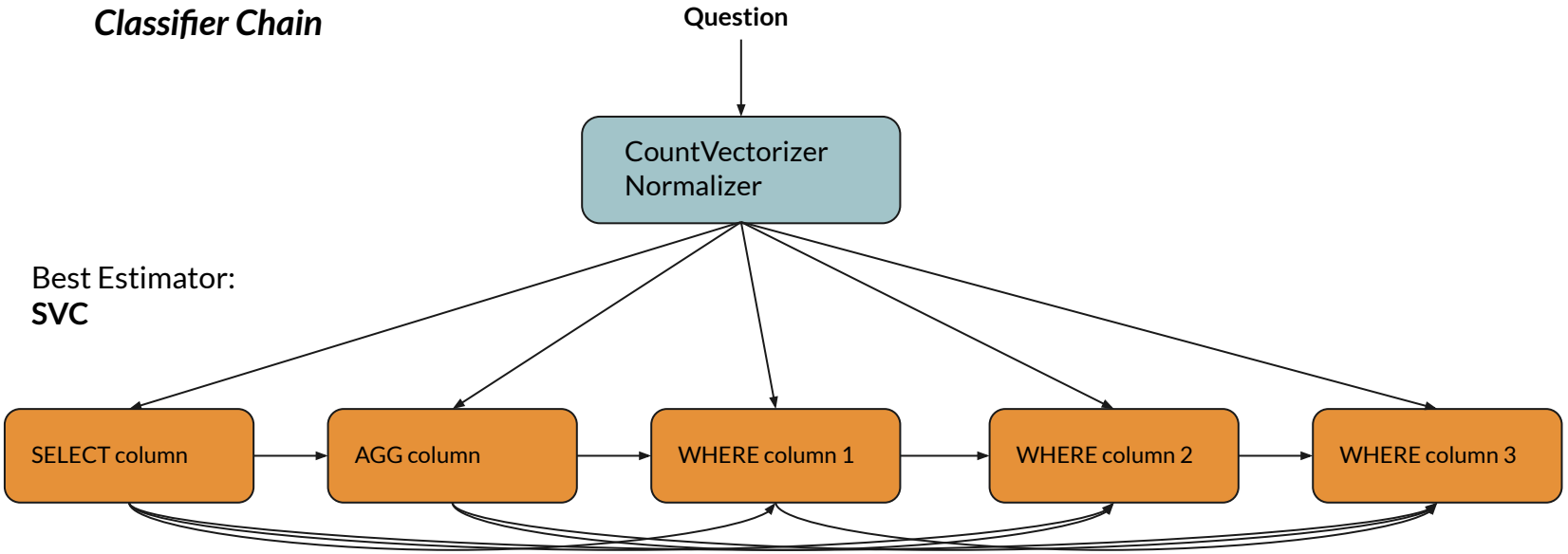
Multi-class multi-output classification problem

Class Imbalances

Data Modeling

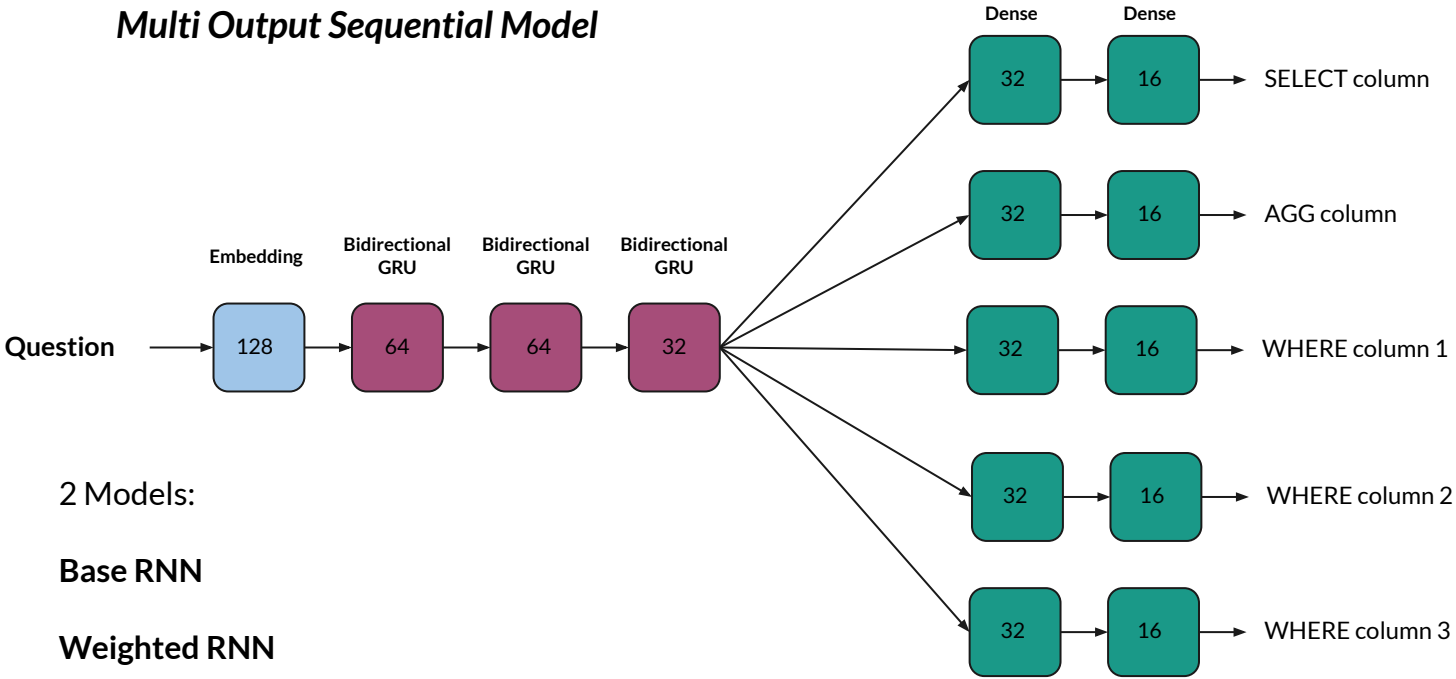


Classifier Chain



RNN Model

Multi Output Sequential Model

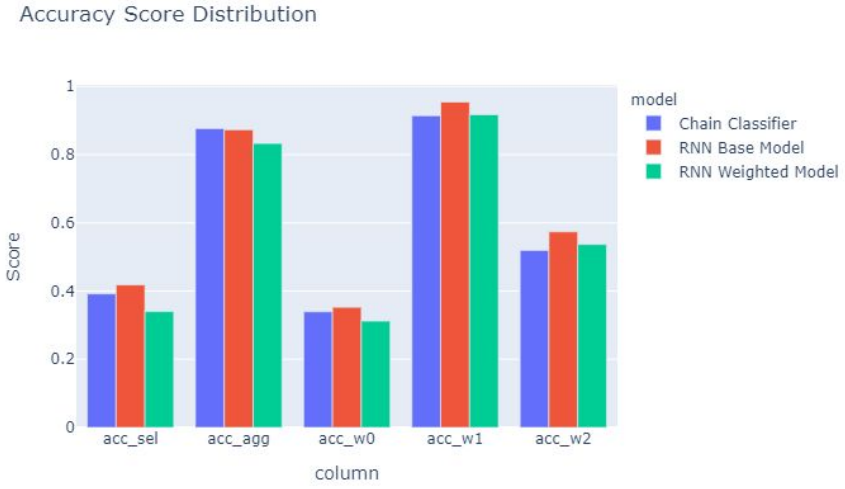


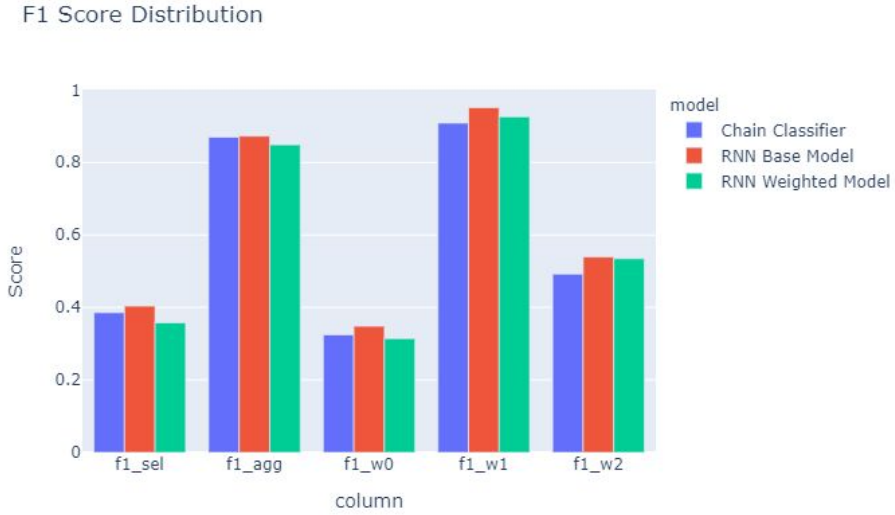
2 Models:

Base RNN

Weighted RNN

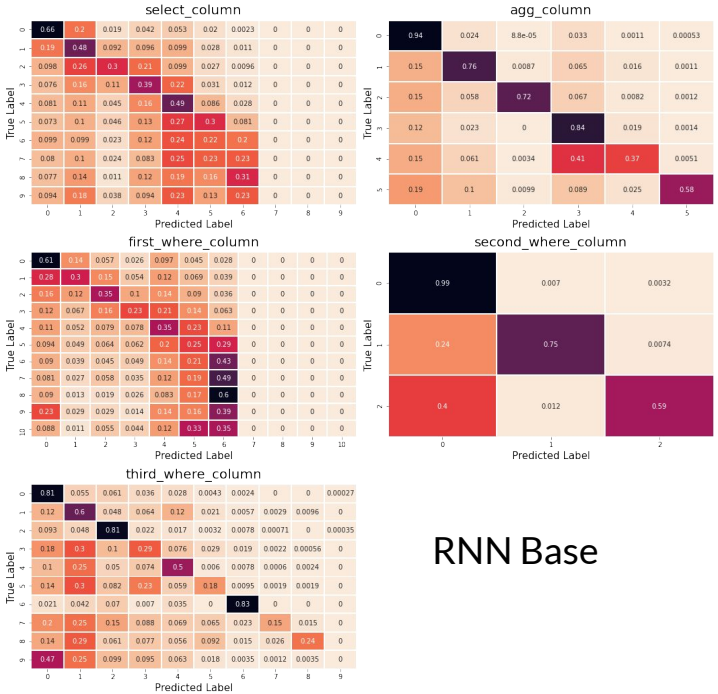
Findings


Accuracy Score Distribution
Score
1
0.8
0.6
0.4
0.2
0
acc_sel
acc_agg
acc_w0
acc_w1
acc_w2
column
model
Chain Classifier
RNN Base Model
RNN Weighted Model


F1 Score Distribution
Score
1
0.8
0.6
0.4
0.2
0
f1_sel
f1_agg
f1_w0
f1_w1
f1_w2
column
model
Chain Classifier
RNN Base Model
RNN Weighted Model

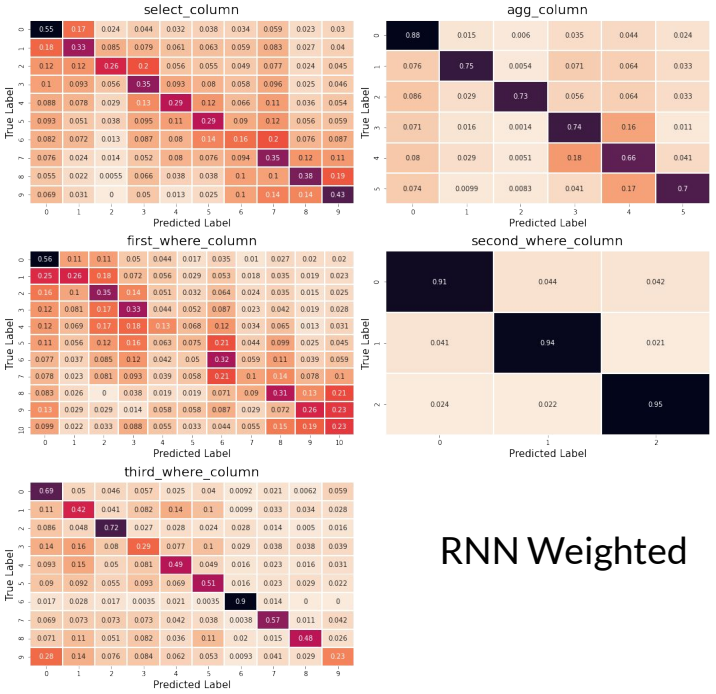
Findings

Query Columns Confusion Matrices



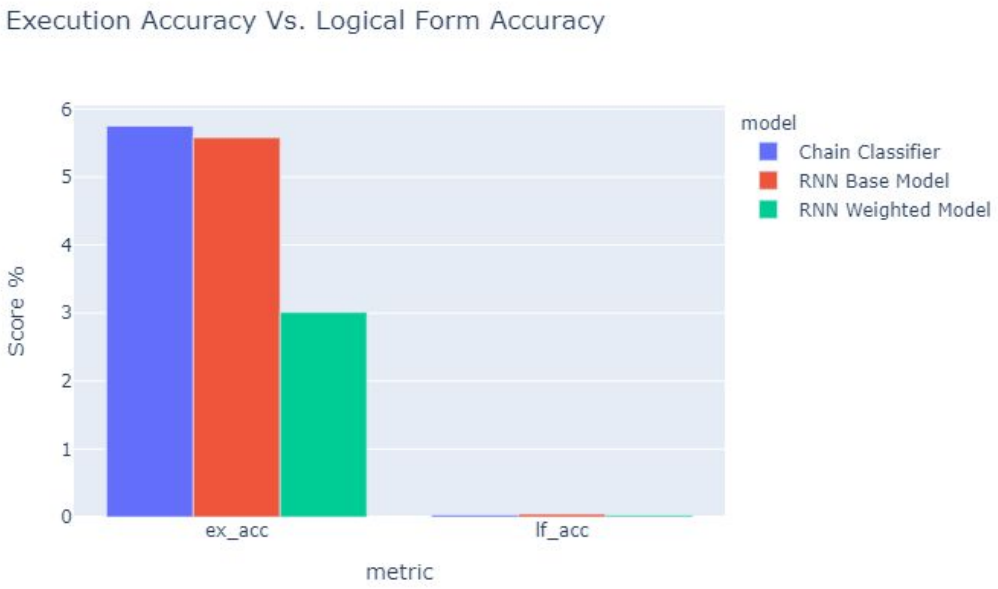
RNN Base

Query Columns Confusion Matrices

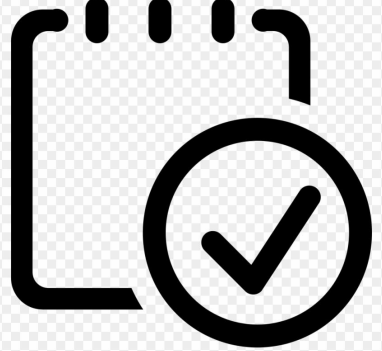


RNN Weighted

Findings


Execution Accuracy Vs. Logical Form Accuracy
Score %
0
1
2
3
4
5
6
ex_acc
lf_acc
metric
model
Chain Classifier
RNN Base Model
RNN Weighted Model

Conclusions



- ➔ The task was challenging
- ➔ The models require high computational power
- ➔ Set up over-simplified the problem (reduced classes and entities)
- ➔ SVC and Classifier Chain worked well

Moving Forward

- ➔ Build an Encoder-Decoder model where outputs are tokens
- ➔ Use different tokenizer
- ➔ Seq2sql model

Questions?