Business Science Problem Framework

How to solve business problems with data science (customer churn problem):

Remember: Businesses want to know the roadmap. Having a plan helps you communicate, clears up uncertainty, and shows stakeholders where they fit into the process. And it can instantly turn you into a leader.

Customer churn refers to the act of customers leaving. This problem is costly, and for large companies' customer churn can cost *R500K per year or more*. These could be subscribers to a software or service or physical customers who choose to shop somewhere else.

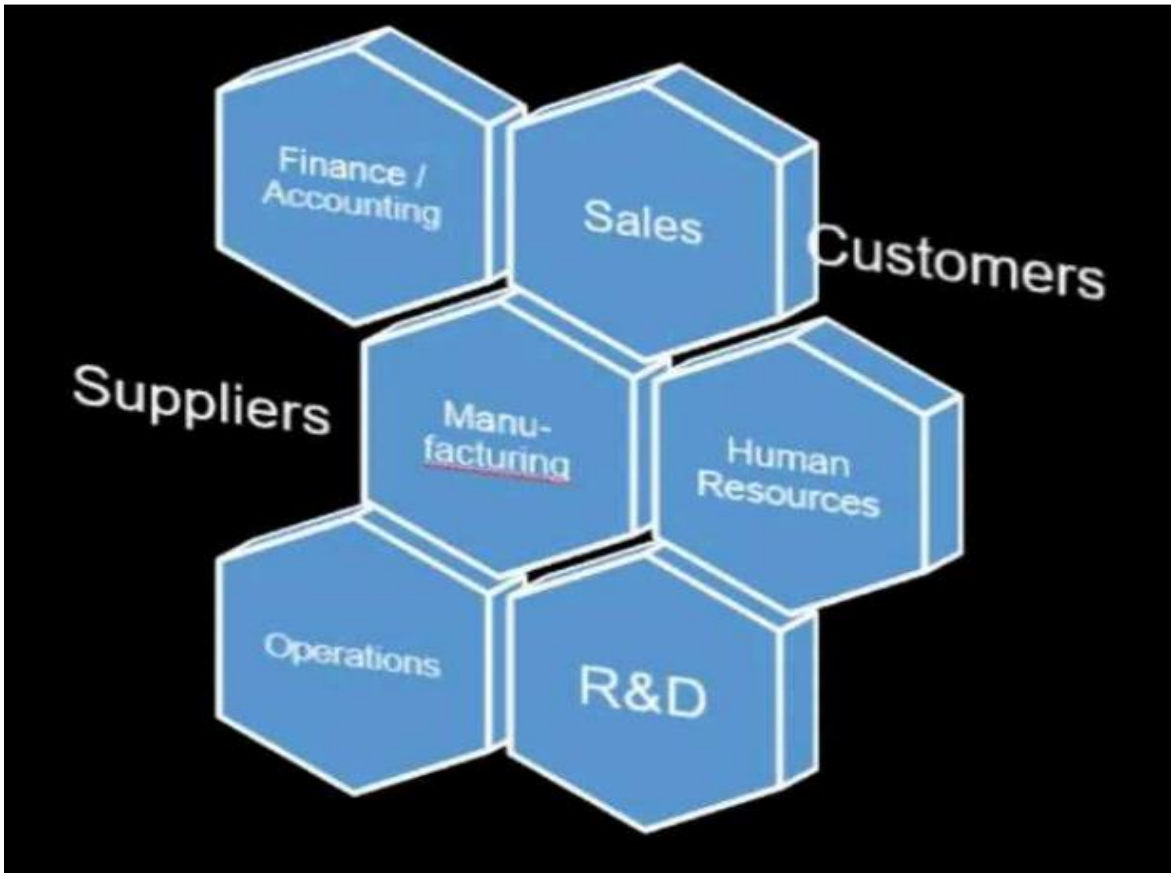
Customer churn often goes undiagnosed. This is because the individual customer impact is usually small, but when aggregated, the effect of churn can be large.

Step 1: View Business as a Machine

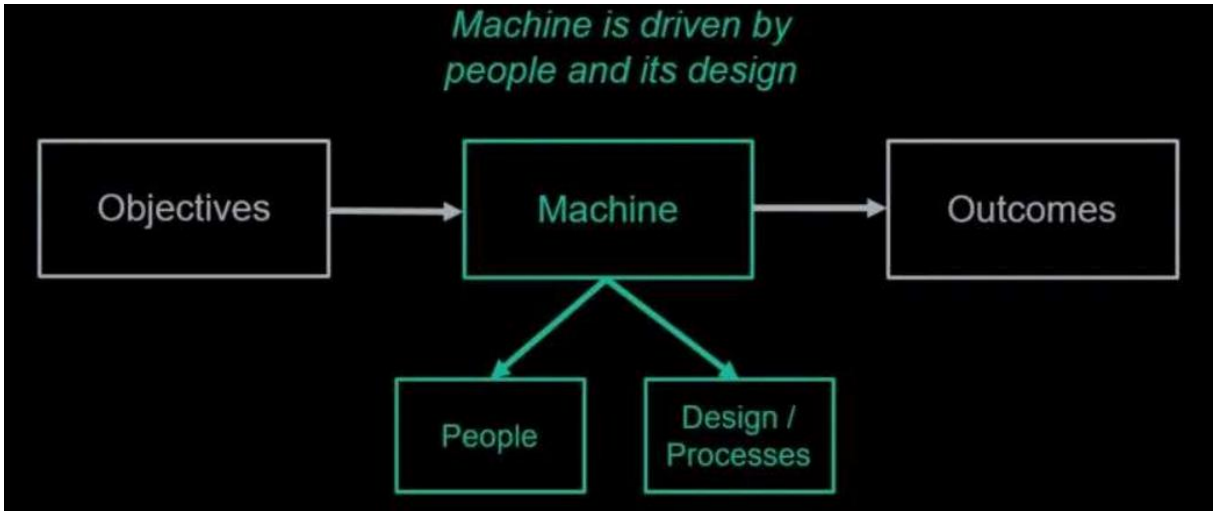
The first step is viewing the business as a machine. This requires the business scientist to:

1. **Isolate business units**
2. **Define objectives**
3. **Collect outcomes**

This involves breaking the business into internal parts (Sales, Accounting, Personnel, etc.) and external parts (customers, suppliers) visualizing the connections.



We then need to visualize this interaction as a machine with goals and outcomes. The goals relate to business objectives, outcomes are what happens. The machine has inner workings, driven by people and processes, the process defines the setup, and the people execute the plan



For example customer churn problem, we make the following assessment:

1. Isolate business units: The interaction occurs between Sales and the Customer.
2. Define objectives: Make customers happy.
3. Collect outcomes: We are slowly losing customers. It's lowering revenue for the organization by R2 million per year

A key aspect of this understanding is the cost of the problem. How is customer loss impacting revenue? R100, R100 000, or R1 000 000?

A good of thumb is that we only want to focus on business problems that are R1 million annually or more.

The higher the cost, the more important it is to solve and easier it is to save the organisation money (ROI)

Step 2: Understand the Drivers

Next, we begin the process of understanding the drivers. The key steps are:

1. **Investigate if objectives are being met**
2. **Synthesize outcomes**
3. **Hypotheses drivers**

Start with the business objective: Customer satisfaction. When customers are happy, they come back. Loss of customers indicates low satisfaction. This could relate to the availability of products, poor customer service, or competition.

Next, we need to synthesize outcomes. In our hypothetical example, customers are leaving for a competitor. In speaking with Sales, several customers have stated "Competition has faster delivery"

The final step is to hypothesize drivers. At this stage. It's critical to meet with subject-matter experts (SMEs). These people in the organisation are close to the process and customers. What are the lead time drivers?

Form a general equation that they help to create:

LeadTime=f(SupplierDelivery,InventoryAvailability,Personnel,SchedulingProcess…)

For example customer churn problem, we make the following assessment:

1. Investigate if objectives are being met: **No, customers are unhappy.**
2. Synthesize outcomes: **Competitor has a faster lead time.**
3. Hypothesize drivers: **Lead time is related to supplier delivery, inventory availability, personnel, and the scheduling process.**

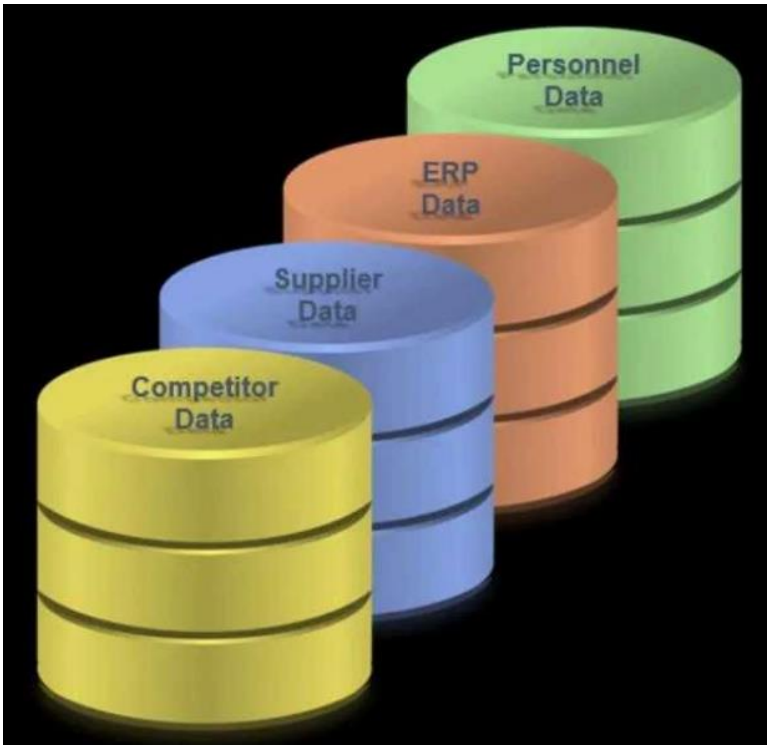
Communication is essential in this stage. As a data scientist, we know the tools well but tools are only useful when we understand the drivers and the business problem. We need to educate ourselves by listening to SMEs.

Step 3: Measure Drivers

Now we begin the process of measuring the drivers. The key steps are:

1. **Collect data**
2. **Develop KPIs**

First, we collect data related to the high-level drivers. This data could be stored in databases or it may need to be collected. We could collect competitor data, supplier data, sales data (enterprise Resource Planning of ERP data), personnel data, and more.



Second, we develop Key Performance Indicators (**KPIs**), which are quantifiable measures that the organisation uses to gauge performance. For our customer churn example:

1. Average Lead Time: 2 weeks, based on customer feedback on competitors.
2. Supplier Average Lead Time: 3 weeks, based on feedback related to our competitor's suppliers.
3. Inventory Availability Percentage: 90% based on where customers are experiencing unmet demand. This data comes from the ERP data comparing sale requests to product availability.
4. Personnel Turnover: 15% based on the industry averages.

Here's what the table looks like with the KPIs added to key process inputs:

KPI Metric	KPI Level	KPI Basis
Avg Lead Time	2 weeks	Competitor lead time
Supplier Avg Lead Time	3 weeks	Competitor supply lead time
Inventory Availability Percentage	90%	Customer sales data
Annual Personnel Turnover in Planning Area	15%	Industry average

There are two key aspects to keep in mind during this step:

1. Collecting data takes time. It may require establishing processes to collect it, but developing strategic data sources becomes a competitive advantage.

2. 2. KPIs require knowledge of customers and industry norms. Data is available outside of your organisation, learn where to access it. Process stakeholders can often direct you where to look.

Step 4: Uncover Problems and Opportunities

In uncovering problems and opportunities we need to complete these steps:

1. **Evaluate performance vs KPIs.**
2. **Highlight potential problem areas**
3. **Review the project for what could have been missed**

For our customer churn example, we review the results from organisational findings against the KPIs to determine where the problem areas exist. We extended the KPI table to include an Actual Value and Conclusion vs the KPI Level:

1. Lead time is 6 weeks compared to the competitor's lead time of 2 weeks.[Concern]
2. Supplier lead time is on par with our competitors [No Concern].
3. Inventory percentage availability is 80% which is too low to maintain a high customer satisfaction level [Concern].
4. Personnel turnover in key areas is zero over the past 12 months [No Concern].

Here's what the KPI table now looks like with actual results and conclusions added:

KPI Metric	KPI Level	Actual Value	Conclusion
Avg Lead Time	2 weeks	6 weeks	Customer churn increasing
Supplier Avg Lead Time	3 weeks	3 weeks	OK
Inventory Availability Percentage	90%	80%	Problem
Annual Personnel Turnover in Planning Area	15%	None	OK

Remember to ask questions and constantly test your assumptions. Talk with Subject Matter Experts to make sure they agree with your findings so far.

Step 5: Encode Decision Making Algorithms

The key parts of this step are:

1. **Develop algorithms to predict and explain the problem**
2. **Optimize decisions to maximize profit**
3. **Use recommendation algorithms to improve decision making**

First, develop algorithms using advanced tools like H2O Automated Machine Learning and LIME for black-box model explanations.

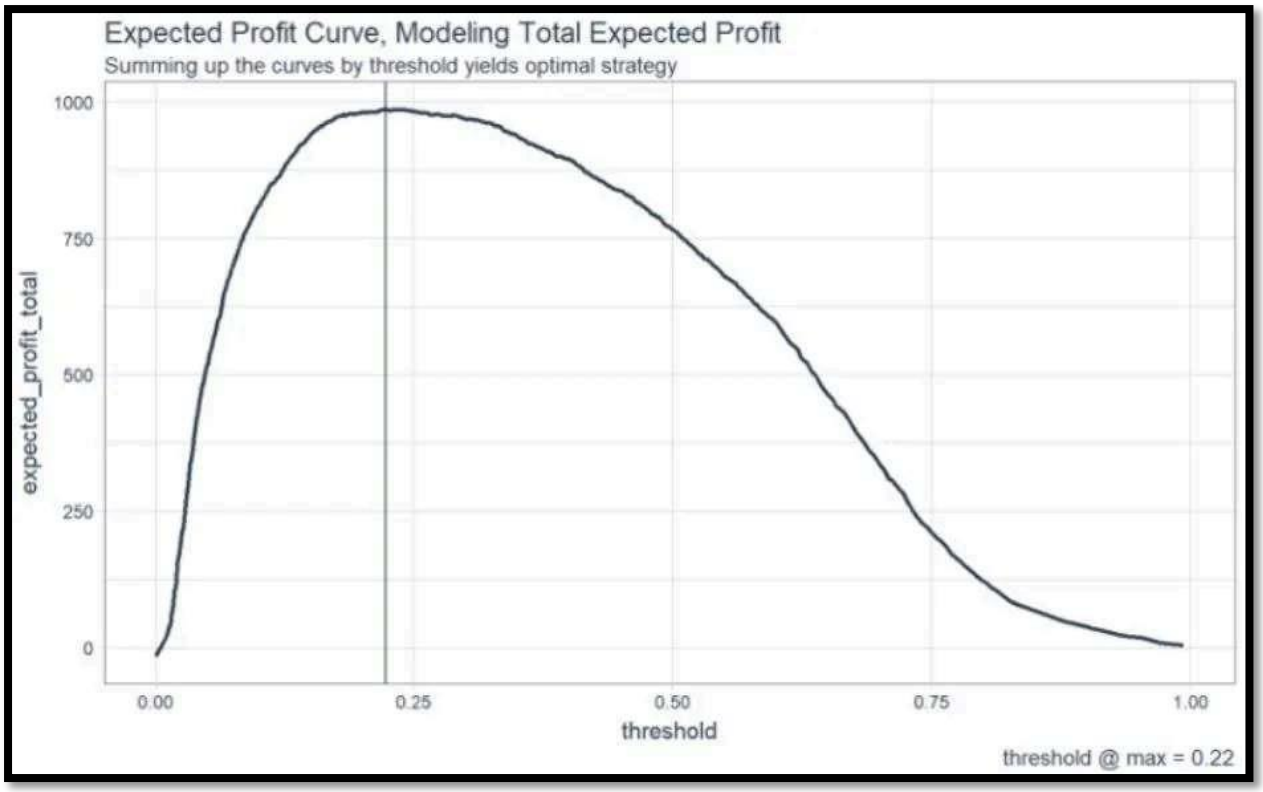
1. **H2O** is a great option because of Automated Machine Learning (**AutoML**). Automated machine learning is fast and develops highly accurate models, saving the data scientist time.
2. 2. **LIME** is used to explain deep learning, random forest, and stacked ensembles, which are traditionally unexplainable. Here is an example of H2O AutoML code to make the models:

```
import h2o
from h2o.automl import H2OAutoML
from lime.lime_tabular import LimeTabularExplainer

# Initialize H2O
h2o.init()

# Convert hr_data_bake_tbl to H2OFrame
hr_data_bake_h2o = h2o.H2OFrame(hr_data_bake_tbl)

# Split the data into train, valid, and test sets
hr_data_split = hr_data_bake_h2o.split_frame(ratios=[0.7, 0.15], seed=1234)

train_h2o = hr_data_split[0]
valid_h2o = hr_data_split[1]
test_h2o = hr_data_split[2]

# Define predictor and response variables
y = "Attrition"
x = [col for col in train_h2o.columns if col != y]

# Train AutoML model
automl_models_h2o = H2OAutoML(max_runtime_secs=200)
automl_models_h2o.train(x=x, y=y, training_frame=train_h2o, validation_frame=valid_h2o, leaderboard_frame=test_h2o)

# Get the leader model
automl_leader = automl_models_h2o.leader

# Explain the model using LIME
explainer = LimeTabularExplainer(train_h2o[:, 1:].as_data_frame(),
                                 mode="regression" if automl_leader.type == "Regression" else "classification",
                                 training_labels=train_h2o[y].as_data_frame().values.flatten(),
                                 feature_names=train_h2o[:, 1:].col_names)
```

Next, optimize decisions to maximize profit.

1. Investigate threshold optimization for binary classification problems.
2. Also, try sensitivity analysis to gauge which features have the largest effect on the probability of the decisions.

Here's what an optimization curve looks like to maximize total expected profit. **Picking the correct threshold is essential. It's the difference between R100 million profit and R75 million or worse.**



The last part is to build recommendation algorithms to improve decision-making. Incorporate feedback from SMEs along with the feature explanations from LIME (or similar feature explanation procedures).

Here's what the code for the recommendation algorithm for **Management's Churn Prevention Strategies** looks like:

```
# Define the conditions and corresponding strategies
conditions = [
    hr_data_raw_tbl['PerformanceRating'] <= 2,
    (hr_data_raw_tbl['TotalWorkingYears'] <= 5) | (hr_data_raw_tbl['YearsAtCompany'] <= 3),
    ((hr_data_raw_tbl['YearsInCurrentRole'] >= 4) | (hr_data_raw_tbl['YearsAtCompany'] >= 7)) &
    (hr_data_raw_tbl['PerformanceRating'] >= 3) & (hr_data_raw_tbl['JobSatisfaction'] == 4),
    (hr_data_raw_tbl['JobInvolvement'] >= 3) & (hr_data_raw_tbl['PerformanceRating'] >= 3) &
    (hr_data_raw_tbl['JobSatisfaction'] >= 3)
]

strategies = [
    "Create Personal Development Program",
    "Promote Training and Formation",
    "seek Mentorship Roles",
    "Seek Leadership Opportunities"
]

# Apply the conditions and strategies to create a new column 'pers_dev_strategy'
hr_data_raw_tbl['pers_dev_strategy'] = np.select(conditions, strategies, default="Retain and Maintain")

# Select the required columns
selected_columns = [
    'PerformanceRating', 'JobInvolvement', 'JobSatisfaction',
    'TotalWorkingYears', 'YearsInCurrentRole', 'YearsAtCompany', 'pers_dev_strategy'
]

# Select the required columns
hr_data_processed_tbl = hr_data_raw_tbl[selected_columns]
```

Step 6: Implement and Measure the Results

Now, it's time to implement the solution and measure results. A powerful tool is a Web App with Dash that can be used to predict churn and recommend management strategies.

This is an example of a web app that implements management strategies:

After the web application is deployed, the results must be measured to show progress. This requires more analysis. We capture outcomes over time and synthesize results.

We measure the effect on KPIs that look like these:

Step 7: Report Financial Impact

If we have done good data science, implemented systematic decision-making, and iterated through problems, correcting along the way, we should now see positive results. Here are the steps:

1. **Measure actual results**
2. **Tie to financial benefits**
3. **Report financial benefit to key stakeholders**

Once the results are understood, we need to show the results as financial benefits.

Here is an example of adding the **financial benefit (net profit and cumulative net profit) over time. This is the ROI that companies need to know to communicate project success:**