

Beuth University of Applied Sciences

# HappiEmployi

## Mental Health Assessment Software for your Employees

### Business Value

Prof. Dr. Sandra Dressler

Teodor Chiaburu

Olena Horyn

Yusif Ifraimov

Silke Meiner

Ilia Ozhmegov

Berlin  
2021

# Table of contents

<b>Table of contents</b>	<b>2</b>
<b>1. Introduction</b>	<b>3</b>
Background (Silke)	3
Description of data product (Teo)	3
<b>2. Market Analysis (Olena)</b>	<b>7</b>
Total Addressable Market	7
Serviceable Available Market	8
Serviceable Obtainable Market	8
Competitors	8
<b>3. Growth Strategies (Yusif)</b>	<b>11</b>
<b>4. Business Models (Yusif)</b>	<b>15</b>
Business Model Canvas (Teo)	17
<b>5. Technical Design and Implementation (Ilia)</b>	<b>18</b>
Data Journey	18
Technology Stack	19
Employee Cost	19
Tech Cost	19
External Data Analysis	20
Data Preprocessing step	20
Feature Selection step	22
Modeling step: Hyperparameter Optimization	24
Modeling step: Validation	26
Modeling step: Testing	27
Conclusion	27
Machine Learning Canvas	28
<b>6. Cash Flow Prediction</b>	<b>29</b>
Cost structure (Olena)	29
Cash Inflow (Silke)	29
Fees HappiEmployi	29
Fees email support	30
Cash Flow (Silke)	31
Investment Capital (Olena)	31
<b>7. Legal structure (Silke)</b>	<b>32</b>
UG & GmbH	32
<b>References</b>	<b>33</b>

# 1. Introduction

## Background (Silke)

For Germany we estimate the GDP (Gross Domestic Product) per working person to be around 80'000€<sup>1</sup>. Employees are absent from work due to health issues for 11 days a year, on average, in Germany<sup>2</sup>. And 17% of the time this is due to a mental health issue<sup>3</sup>.

In a company of 150 employees we lose around 90'000€ a year to mental health.

Now this is not good. And we might add that in IT companies we have a higher value creation per employee than the estimated average 80'000 per year. We might add that in IT our employees are under a lot of pressure and we might have seen cases of burnout and depression among our colleagues.

This comprises cost for our employees that suffer from mental health issues. We'd like to focus attention also on the cost of losing an employee, and his or her knowledge of the company and our processes, their network among our clients. This can not quickly be replaced, sometimes not at all.

Replacing an employee can be costly, directly through paying an recruiting agency. For a manager or technical lead this might be 10 to 15K€. We might add in the time our management and HR personnel spends in the recruiting process even with only 5 interviews. And we might consider the drop in performance of the team while their manager / tech lead is still to be found or still getting to know people and procedures and catching up with their tasks. And even when all this is accomplished, there is the risk that the new hire might not fit the company culture or the specific team to work with.

However, we might also argue differently, that we don't look at our employees in this statistically and accumulated way, but as individuals and as human beings. And some of them need our help. Finding out who is in need of help due to a mental health issue can be difficult. It becomes easy with HappiEmployi!

Speaking of KPIs our product will reduce the days employees of sick leave and reduce the employee churn rate.

## Description of data product (Teo)

Both physical and mental health are a critical factor in having a successful career and a joyful life. A happy and healthy employee will invariably contribute to the progress of the company they are working for. In the context of the current impediments that the pandemic is posing on the working market (offices closed, online schooling, financial cutdowns), people's

---

<sup>1</sup> <https://de.statista.com/statistik/daten/studie/1252/umfrage/entwicklung-des-bruttoinlandsprodukts-je-einwohner-seit-1991/>

<sup>2</sup> <https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Qualitaet-Arbeit/Dimension-2/krankenstand.html>

<sup>3</sup> <https://de.statista.com/statistik/daten/studie/77239/umfrage/krankheit---hauptursachen-fuer-arbeitsunfaehigkeit/>

mental well-being has become a central issue of debate. On a brighter note, the past year has been very fruitful for e-health services. Now more than ever, there is a dire need for good-quality medical assistance delivered online. It is this business sector that we intend to dive into, by offering our help to employers that want to tend to their employees' mental health in these difficult times and beyond.

More precisely, our product - HappiEmployi - is a web application consisting of an elaborate query with questions regarding your general mental health and work-life-balance (see <https://happiemployi.000webhostapp.com> ). The question types range from multiple choice to assigning scores that assess how you feel towards different aspects of your company's organization or even writing short texts describing your state (see *Figure 1.1*).

Do you work remotely (outside of an office) at least 50% of the time?

☐ Yes

☒ No

Does your employer provide mental health benefits?

☐ Yes

☐ No

☒ Don't know

Has your employer ever discussed mental health as part of an employee wellness program?

☐ Yes

☒ No

☐ Don't know

How easy is it for you to take medical leave for a mental health condition?

☐ Very difficult

☒ Somewhat difficult

☐ Somewhat easy

☐ Very easy

☐ Don't know

*Figure 1.1:* Snippet from the query (see [link](#) )

The example form can be accessed for free on our website, meaning that employers would only need to send their employees our link, either via *slack*, which we offer an integration for, or delegate the task to their HR (for larger companies we also offer *HRIS* - Human Resource Information System - integration). After completion, the employer - who also has a separate admin page in the account he created on our site - will simply have to submit to us the answers they gathered (see *Figure 1.2*).

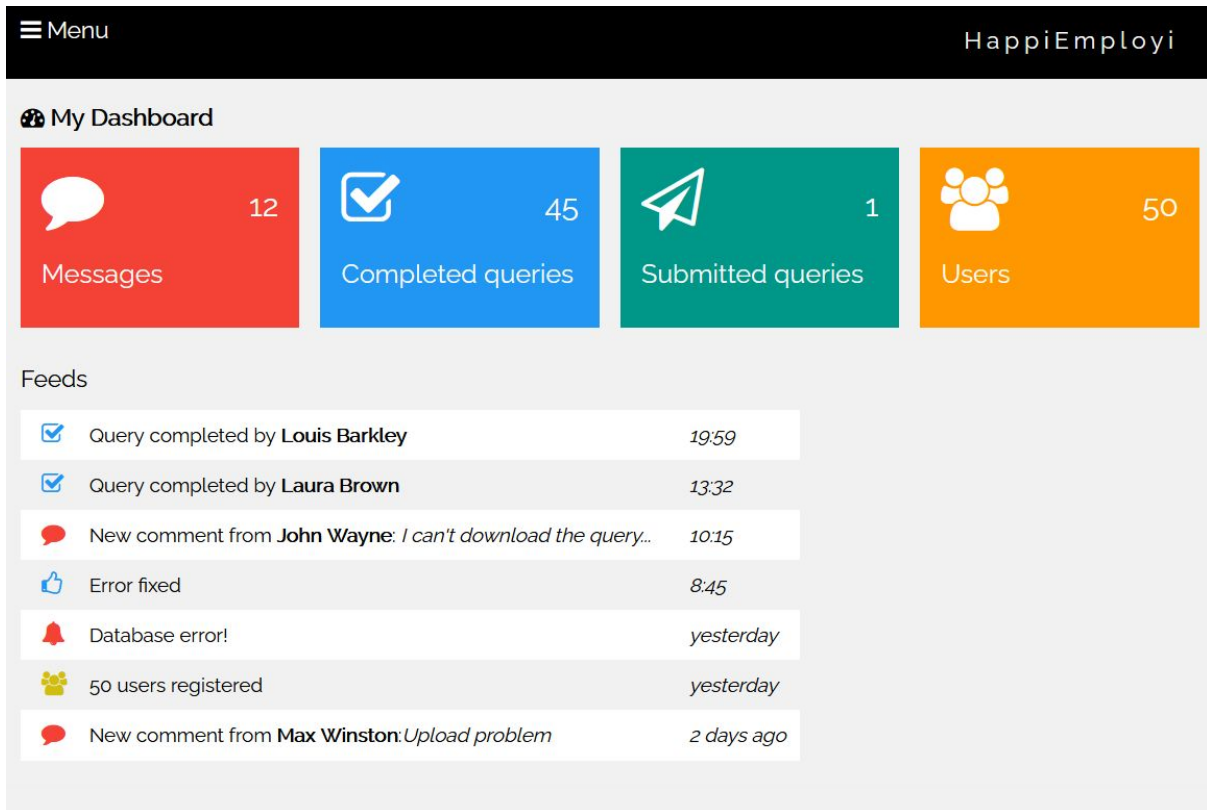


Figure 1.2: Employer's admin page (see [link](#) )

Our team of data scientists will then apply the most state-of-the-art machine learning algorithms in order to precariously identify potential mental health risks among the employees. Then, in the case of suspicious cases, in collaboration with our experienced psychologist, we will come up with a detailed diagnosis for each individual that appears to be suffering from some form of mental problem. The results are then sent back to the employer, which gets notified per email and can open the statistics on his admin page. Depending on which pricing plan was chosen, the employer and their team may also benefit from one-on-one coaching from our side (see *Figure 1.3* and *Figure 1.4*).

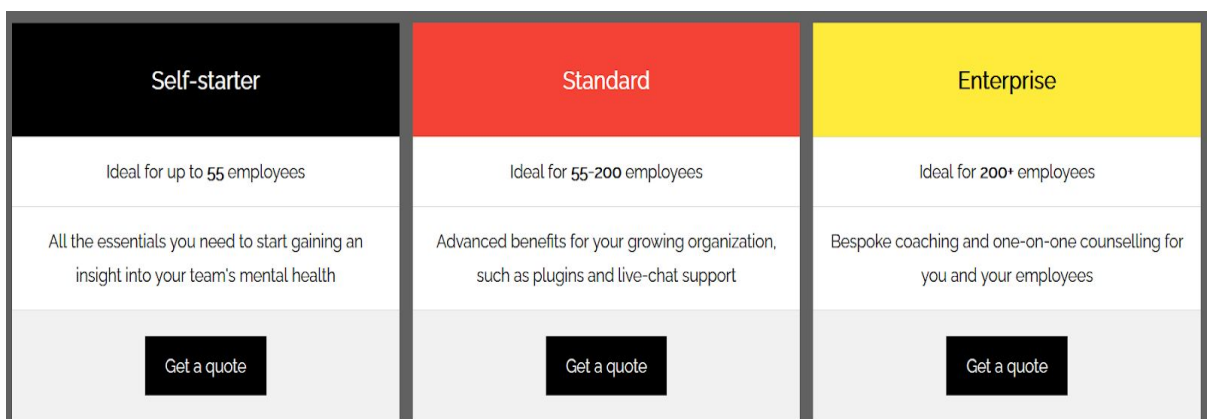


Figure 1.3: The 3 packages the employer can choose from

	Self-starter	Standard	Enterprise
Slack integration	✓	✓	✓
Data extract in CSV, PDF, Excel	✓	✓	✓
Chrome and Outlook plugins	✗	✓	✓
HRIS integrations	limited	✓	✓
GDPR compliant	✓	✓	✓
ISO27001 certified	✓	✓	✓
Coaching	group	group	one-on-one
Support	email	email + live chat	email + live chat

*Figure 1.4: Comparison of the 3 packages from Figure 1.3*

## 2. Market Analysis (Olena)

The Mental Health software market has been growing and developing rapidly in the past few years. It is estimated that the global Mental Health software market was worth 1.36 billion in 2020 and is expected to reach 2.6 billion by 2025<sup>4</sup>.

### Total Addressable Market

The Total Addressable Market (TAM), also referred to as Total Available Market, is the overall revenue opportunity that is available to a product or service if 100% market share was achieved. Using this TAM metric will help us to determine the level of effort and funding that we should put into our new business.

There are three methods used to calculate the total addressable market. They include: Top down approach, Bottom up and Value theory.

#### Top down

The top-down analysis follows a process of elimination that starts by taking a large population of a known size that comprises the target market and using it to narrow down to a specific market segment.

There are about 3.5 million companies in Germany that are medium-sized enterprises. This can be a size of our target market when we use this approach.

#### Bottom up

Bottom-up analysis uses data on the current pricing and usage of a product. We sell a B2B software solution that works best for mid-size companies with a 50+ employees, we can calculate TAM using the following formula:

$$\text{TAM} = (\text{Annual Contract Value}) \times (\# \text{ of possible Accounts})$$

Since our business strategy is to keep our product free of charge for the first year while we collect enough data, this TAM approach is not suitable for us.

#### Value theory

Value theory relies on an estimate of the value provided to customers by the product and how much of that value can be reflected in the product pricing. A company estimates how much value it can add and why it should capture this value through the product pricing. Value theory is used to calculate TAM when a company is introducing new products into the market or cross-selling certain products to existing customers, therefore we cannot use this strategy yet at such an early stage of our company. We could possibly use it in the future when introducing new products or services to our existing clients.

We will use a top down approach in evaluating TAM.

---

<sup>4</sup> <https://www.marketdataforecast.com/market-reports/mental-health-software-market>

## Serviceable Available Market

Serviceable Available Market or Served Available Market (SAM) is the target addressable market that is served by a company's products or services. SAM is the segment of the Total Addressable Market.

Our TAM consists of 3.5 million companies in Germany. If we narrow it down to only IT Software and Hardware companies - we'll have more than 94,000 of potential customers just in Germany alone<sup>5</sup>.

## Serviceable Obtainable Market

The Serviceable Obtainable Market, or SOM for short, is the portion of the served available market a company can realistically capture. SOM shows what is realistically being targeted to be captured by the business using the following formula:

SOM = % of SAM that can be realistically obtained

If we only have 5% share of all of the Serviceable Available Market in Germany, we can obtain 4,700 potential customers.

$$\text{SOM} = 5\% * 94,000 = 4,700$$

And if with time we increase the share of our Serviceable Available Market to 10%, we now will be looking at 9,400 potential customers in Germany alone.

$$\text{SOM} = 10\% * 94,000 = 9,400$$

Also, with time we can expand our TAM to the global market, where we would have even more opportunities to grow and to attain more customers worldwide.

## Competitors

The field of Mental Health software is highly competitive and there are many software solutions available. This is a list of our main competitors and what they do.

### SimplePractice

SimplePractice is a practice management platform, made for small business owners in the health & wellness space. Their services include a fully paperless intake process, custom notes & forms, free appointment reminders (SMS, email, & voice), mobile app, e-claim filing, a client portal, billing & invoicing, and the newest feature: Telehealth.

---

<sup>5</sup> <https://www.statista.com/statistics/462234/it-industry-number-of-companies-germany/>



### **TherapyNotes**

TherapyNotes is an online practice management and Electronic Health Records (EHR) software for behavioral health practices of all sizes. They are a platform for mental health professionals, offering the software and customer support. They offer a streamlined scheduling, notes, billing, and a custom client portal and support.

### **CareLogic EHR**

CareLogic services behavioral health and human service providers with tools designed to improve outcomes and demonstrate performance.

### **Psyomics**

Psyomics is developing software to support healthcare practitioners by providing a comprehensive picture of a patient's mental wellbeing. Co-developed by psychiatrists and built on evidence-based mental health research, their digital diagnostic tools can be used to assess and triage patients at all levels, primarily in the UK.

### **Braive**

Braive – used by psychologists for iCBT (internet-Based Cognitive Behavioural Therapy) to treat patients facing mental health challenges.

### **Unmind**

Unmind is an app with the aim of promoting good mental health and wellbeing at the workplace. Unmind offers a wellbeing platform, delivering 'clinically-backed' content and support to help workers. Features include a mood diary, world-class exercises for stress, focus and sleep, as well as learning stories, a chatbot, daily challenges and tasks.

### **Moodpath**

Moodpath helps people recognize the symptoms of depression by answering daily questions over a two-week period. During that period, their app tracks the user's psychological, emotional and physical health and based on the answers, it creates a personalized mental health assessment that users can use as a starting point to talk with therapists, psychologists, and other healthcare professionals.

### **Bluecall**

BlueCall is a Stockholm-based startup offering instant and anonymous therapeutic conversations through an app. Primarily offered to employers so they can create the right conditions for their employees to increase wellbeing, BlueCall starts with a few questions and then matches employees with conversational support by volunteers, mentors, and certified psychologists who are suited to their particular needs. It operates in Sweden only for now.

### **Spill**

Spill is a message-based therapy app for individuals and organizations. Users are paired with one of Spill's qualified counselors who provide support, guidance, and exercises, whom they can contact whenever they need to talk.

**BetterHelp**

BetterHelp is an online counseling platform. They help individuals who face life's challenges by providing access to licensed therapists.

**Why are we different?**

What makes us different is that our software is designed and tailored to employers - mostly mid-sized enterprises with 50+ employees. Employers (companies) are our main clients and they will use our software to evaluate the mental health of their employees (e.i. if the employee at the high or low risk of developing a mental health illness/issues).

When most of our competitors offer Electronic Health Records software for mental health professionals or products tailored toward individuals, we offer an easy and instant solution for companies to help them evaluate their employees mental health risks. At this point we offer one of a kind solution for mid-sized enterprises.

### 3. Growth Strategies (Yusif)

#### Product development and diversification

The development of any Data Product requires a lot of time and effort. However, that is usually worth it as the product brings the competitive advantage that will be hard to repeat due to time required to prepare such models. Illustrated below is our Product Development and Growth strategy.

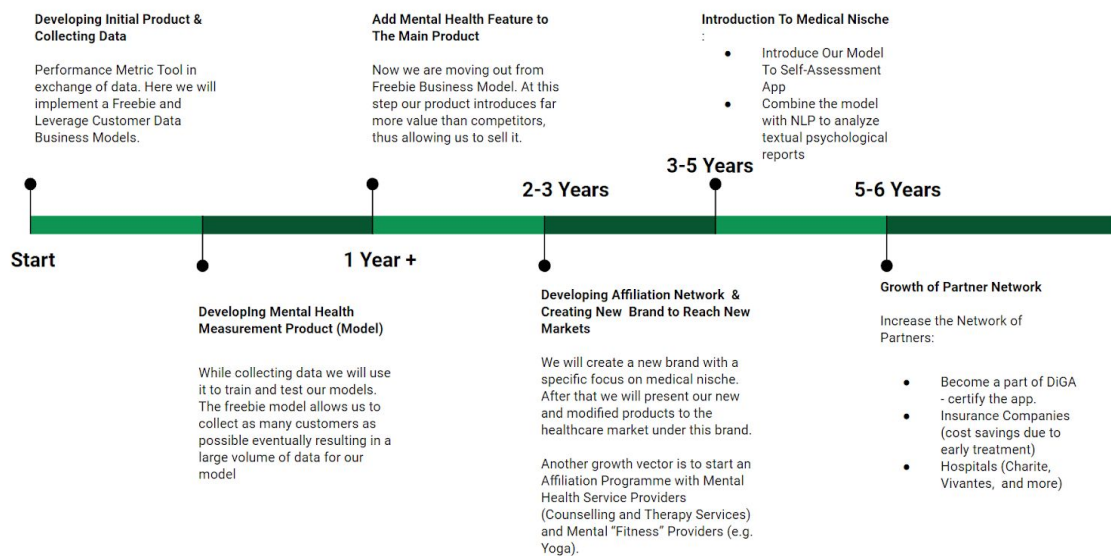


Figure 3.1: Growth Strategy Pipeline

## Product Development

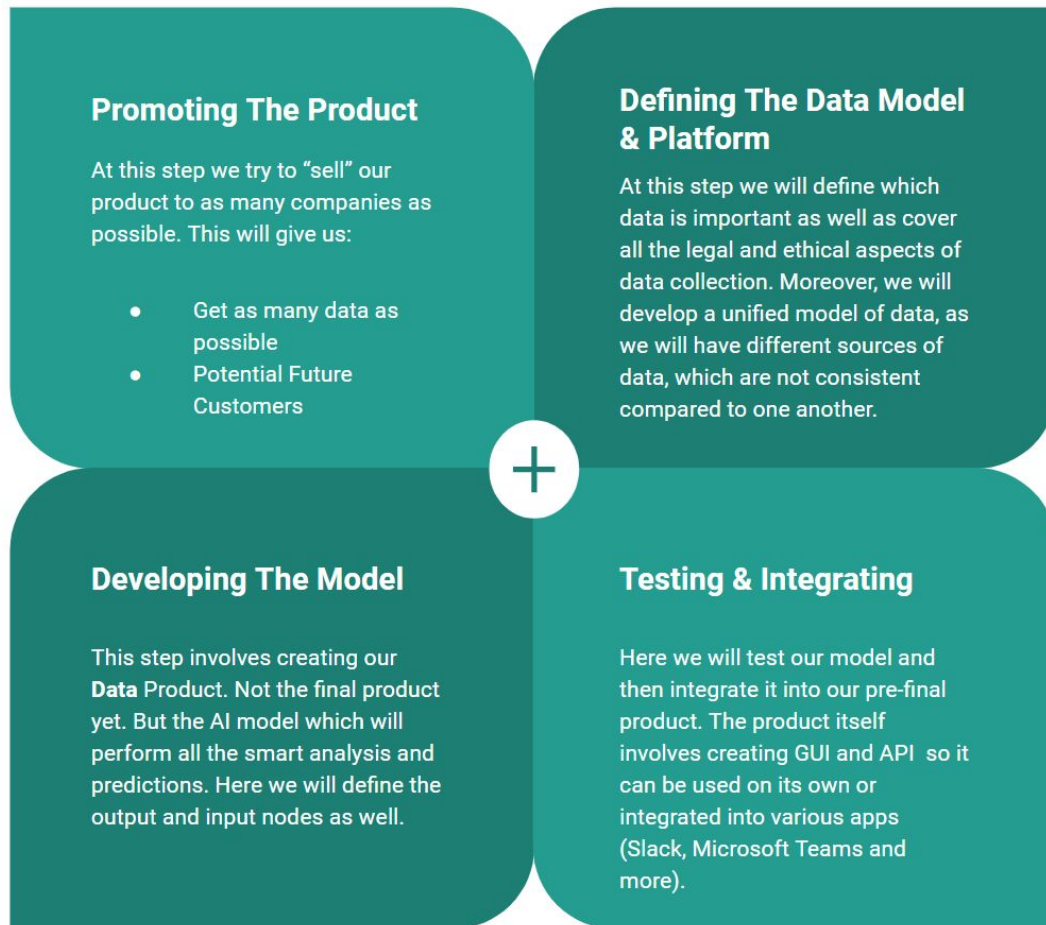


Figure 3.2: Main Product Development Canvas (Top-Left first stage, Bottom-Right last stage)

It is hard to build a Data Product without any data at hand, especially in a developing market where the good quality and large enough data is not available. Because of that we decided to build our own strategy to gather data which will help us in development of our product. The strategy is built around the “Freemium” Business Model. The idea is that we will develop our first product - Performance Metric Tool (which will be developed quite fast) for free. Since there are existing offerings like that in the market we will penetrate by offering the product for free. This in turn should help us build the potential customer base and most importantly collect the data. Projecting this onto Ansoff Diagram we are moving towards Market Penetration Strategy at this point.

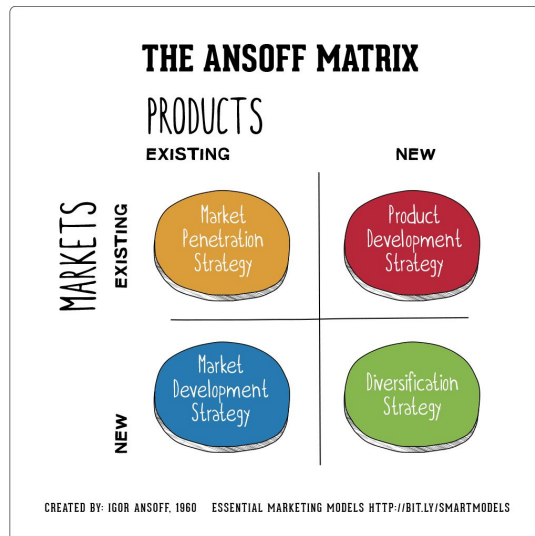


Figure 3.3: The Ansoff Matrix

### Product Diversification

While the first year we will penetrate the market with our “Freemium” Business Model after our main Mental Health Prediction AI Engine will be developed we plan to integrate it into our SaaS, thus diversifying the product. Since the market we penetrate already exists we are moving towards Product Development Strategy according to the Ansoff Matrix. Because at this stage our product introduces more value to our customer, it will mean that the customer will be motivated to pay for the product which does not have an alternative on the market. Thus we move away from the “Freemium” Business Model at this stage.

### Affiliation & Introduction of Diversified Services

After gaining sufficient customer base we plan to run an affiliate programme with companies in the similar market but offering different products. Since our solution is at the beginning of the pipeline (early detection & self-mental diagnosis) we are not engaged at steps happening at the end of the pipeline - employees looking for mental help. However, we can instantly solve this problem by partnering with such service providers (for example Blue Call). The idea is that by bringing the employees in need of mental help to mental help providers we gain certain incentive and on the other side provide a smooth experience to our customers by helping them to connect with appropriate service providers. This creates a win-win situation where customer’s satisfaction increases and we create an additional revenue stream.

### Penetrating new markets, creating a new brand and diversifying the product further

Our AI enhanced SaaS will develop over time providing a better overall performance and precision, as the product matures we plan to introduce it to the new markets. One of which will be the Healthcare Market. The product we propose will be focused on triaging the patients on the way to their mental therapist. A new branding will be created for that to provide a customer a better structural understanding of our offerings. Due to the current pandemic, which we expect to last for another year or two there have been a lot of cases of

Mental Health Issues. The facilities and labor providing mental help are limited in their capacity, thus we expect that there will be a great demand in triaging the patients (giving certain scores to the patients which will help to define how urgent is the case, and whether it should be solved through a direct contact or the patient can proceed with self treatment and very limited visits to the therapist). From a technical point of view we are not diversifying our main product significantly at this stage, but we are penetrating a completely new market - moving towards Market Development Strategy.

**Increasing market reach, introducing new revenue streams and getting DiGA certification.**

Becoming a part of DiGA should introduce an additional revenue stream as our app can be prescribed by doctors and insurance companies or some insurance packages can cover the costs. On the other side insurance companies will benefit from cost savings induced by early problem detection of the mental issues (it is easier and cheaper to treat those at early stages). Eventually, this creates a win-win situation for both sides (patients and insurers). Moreover, being listed in the DiGA repository should increase the market reach for us and promote our product. All in all, we are creating a market where we deliver the value to both parties - insurance companies benefit from cost savings and where patients benefit from getting the treatment on time.

## 4. Business Models (Yusif)

### What are our main business models?

Business models define the whole process within our company. Their unique combination gives us a competitive advantage over our competitors through ingenious cost saving schemes, and tricky and not instantly obvious revenue streams. From the sentence above it becomes clear that we plan to use several of them. While you can get an overall overview of our Business Model in Business Model Canvas, here you will get a projection of our business models onto patterns developed by BMI Lab - [Business Model Navigator](#). This should render down the whole picture into more comprehensible chunks. The models will not be introduced simultaneously, the overall business model will become sophisticated over time. The models below are put into chronological order:

### Reverse Engineering

As mentioned above the product will be offered for free is a Performance Metric & Survey Collection Tool which already exists on the market. The idea of the “Reverse Engineering” Business model is that we save a lot of costs on Research & Development as well as on Design simply by obtaining a competitor's product, taking it apart, and using this information to produce a similar or compatible product. Since no huge investment in research or development is necessary, our alternative products can be offered at a lower price in our case for **free**.

### Freemium

As we are an IT Startup, which provides a SaaS the economy of scale applies to us from the early stages. Large customer base could significantly reduce the costs of the product as the cost will be spread over a larger number of customers (sorry for tautology). In other words we can say that marginal cost of product per additional customer is extremely low. Thus, scaling our customer base at an extreme pace is the key to our success. In order to achieve that we will offer our base offering free of charge, and later as our main product will be developed we will offer it for “premium” cost. This model is called “Freemium”, and the free offering is able to attract the highest volume of customers possible for the company at initial stages.

### Leverage Customer Data

One of key business models is so called “**Leverage Customer Data**” (see. [Leverage Customer Data Business Model Pattern](#)). Since our main value comes from our Data Product integrated into SaaS, the Data itself is a key component in delivering the value to the customer. The data we collect is used to train and test our models. It used both in our base and premium offerings. More details about the data flow and usage is provided in the technical part of this report.

## **Affiliation**

Affiliation should allow us to deliver more value to our customers. Since the mental health issues involve treatment and therapies, our partners and customers may benefit if we bring them together. We plan to use a pay-per-sale model where we get a money incentive each time one of our partners performs a sale (counselling services, yoga and mindfulness sessions). The partners get access to an additional customer flow where the customers get a value of easy access to those services as well as support and best matching for their specific needs (based on the data we collected so far).

## **Solution Provider**

This model should become an end-point in one of the markets we penetrate, specifically HR Solutions/Operating Systems. The eventual idea of following this business model is to offer coverage of services in one particular domain (HR Systems). Most importantly this will allow us to get a broader insight into customer's habit which in turn can be used to improve our products and services. This business model has a close relation to "Leverage Customer Data" business model.

## **Make More of It**

The business models will change with the development of the company. Building a working Data Product requires a lot of time and effort. We understand this and the fact that we should take the most out of our investments, thus after a year of development and eventual release of our final product we plan to follow the "**Make More of It**" business model. Our plan is to introduce our product into a different market, specifically - healthcare. The idea is to build an app based on our existing Data Product & SaaS (or Know Hows) which will help to perform a psychological self-assessment for a user. Eventually this should help to triage the patients based on the severity of their mental condition. Considering the current pandemic the load on psychologists has increased significantly, meaning that triaging the patients became more crucial than ever. Since we are utilizing the existing know-hows but in a new domain/market, this will allow us to save costs and time during the development process. We are "making more" out of what we already have.

## **Digitalization**

This pattern relies on the ability to turn existing products or services into digital variants, and thus offer advantages over tangible products, for instance, easier and faster distribution. This model will be followed as soon as we start developing our automated reporting system specifically dedicated to mental health records. Efficiency and multiplication by means of digitization brings additional value to the doctors in terms of time and money savings. The time savings on the other side brings a value to the patients as they will have more time to communicate with the doctors/therapists.



# Business Model Canvas (Teo)

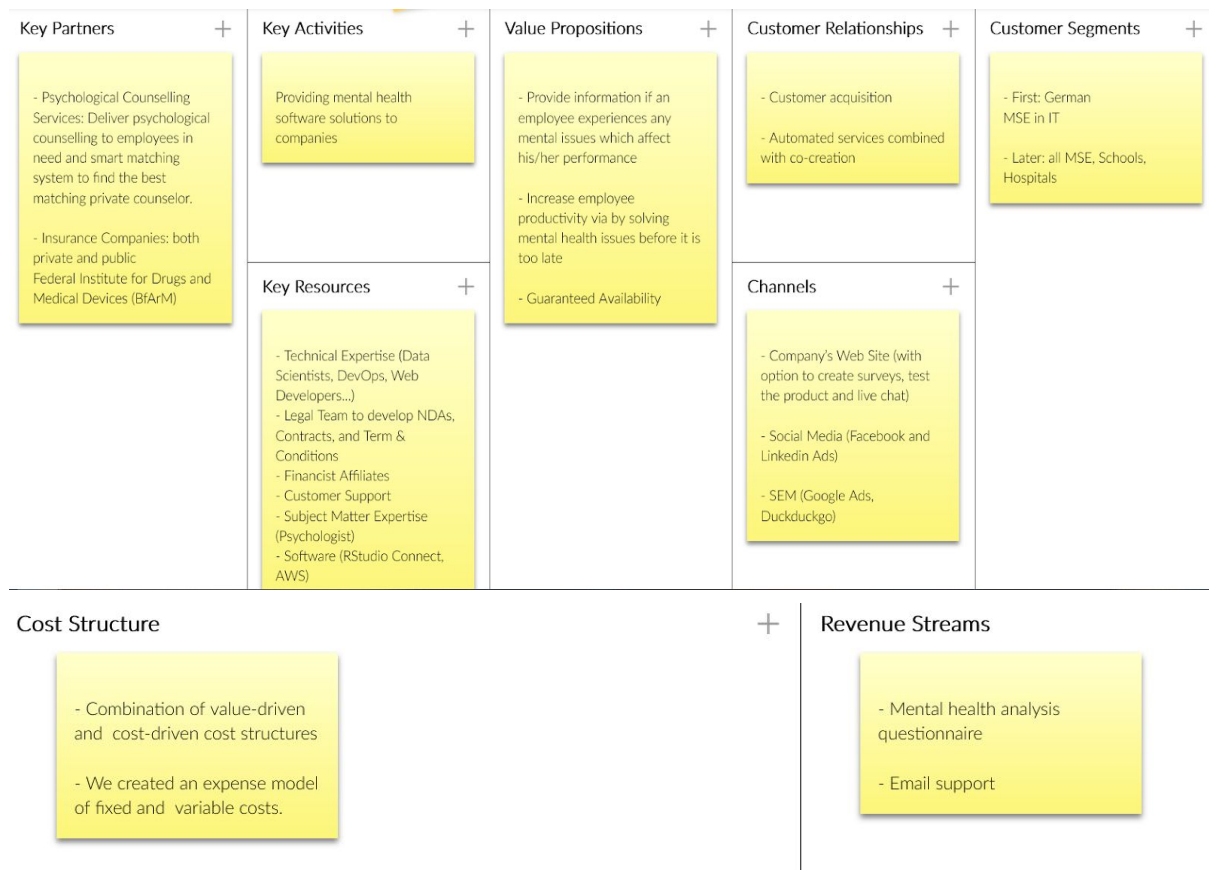


Figure 4.1: Business Model Canvas

## 5. Technical Design and Implementation (Ilia)

Our web-application requires an explanation regarding our most challenging part - Data Product. The rest parts of it are standard so we will focus only on Data Product Parts.

### Data Journey

Data Journey can be divided into two stages according to the growth strategy. The first one covers the data acquisition stage whereas the second one is an iterative process that includes exploratory data analysis, model analysis, model tuning and data transforming.

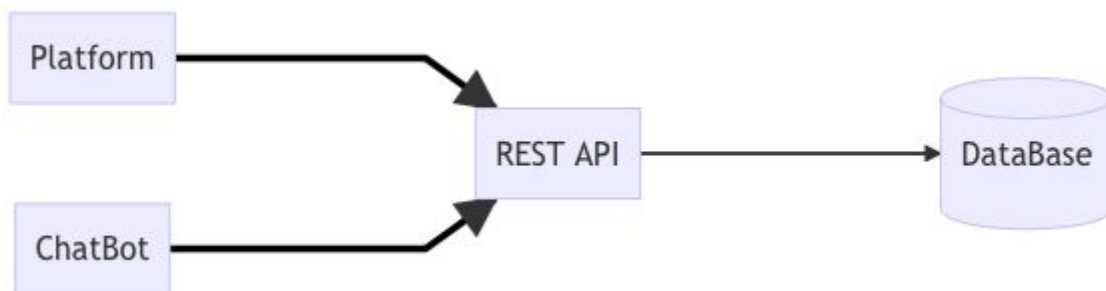


Figure 5.1: Data journey during the first stage

In the diagram above can be seen the data journey of the first stage.

- **Platform** is a website where employees can fill in data.
- **ChatBot** is a simple RASA chat assistant that allows an employee to fill in questionnaires via corporate chat (e.g. Slack, Rocketchat).
- **REST API** is an API that handles requests and puts them into our AWS DataBase.
- **DataBase** is a database where all our data is kept until further steps are taken (PostgreSQL).

Here can be seen our REST API based Handler solution that processes data simultaneously from Platform and ChatBot. ChatBot is supposed to use Python/RASA technology meanwhile Platform and REST API Handler should use standard Kotlin solutions. Machine relative architecture is clear and does not require any detailed explanation. It is expected to be implemented before collecting data from our clients that is why it has a higher priority than the next stage.

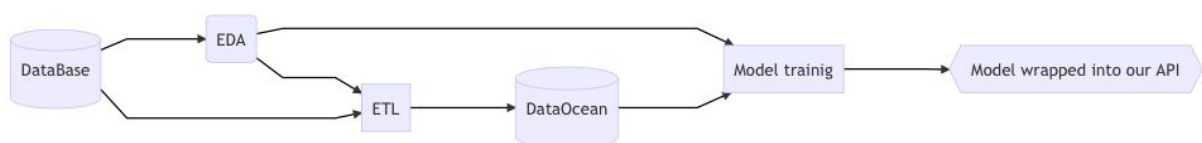


Figure 5.2: Data journey during the second stage

The second diagram explains the continuous iterative process of Data Journey which is the final target to create/update our model access to which will be done via REST interface as the result of the provided data.

- **EDA** - explanation data analysis, includes also a model search. It is expected to update ETL and the Model manually.

- **ETL** - Extract, Transform and Load to another database that we call "DataOcean".
- **DataOcean** - a database where we keep data prepared for Model training (Redshift).
- **Model Training** - training model based on the entire data and the "recipe" provided from the EDA block.

## Technology Stack

Taking into account the previous section and general needs we show you the most important technologies we are going to use.

Data Science	R (tidy*), Python (NLP frameworks)
Data/DevOps Engineering	SQL (PostgreSQL, Redshift), AWS
Backend Engineering	Java/Kotling/Scala
Web (Front end)	HTML, CSS, JS

*Table 5.1: Technology stack*

## Employee Cost

Employee	Requirements	Type	Salary
Data Scientist (me)	R (tidyverse, tidymodels, recipes), Python (NLTK, CoreNLP), REST API, Linux, AWS, SQL (Redshift, PostgreSQL)	full-time	100k EUR
Data/DevOps Engineer	Kafka/Spark, CI/CD, SQL (Redshift, PostgreSQL), Linux, Python, AWS, REST API	full-time	60-70k EUR
Backend Engineer	Kotling/Scala, AWS Linux, REST API	full-time	60-70k EUR
Web Developer	HTML 5, CSS 3 (e.g. Bootstrap), JavaScript/TypeScript	part-time	60k EUR

*Table 5.2: Employee cost*

## Tech Cost

In table 5.3 you can see tech cost. Hosting is needed for Staging, Production, Handling, Platform, ChatBot and probably more. The cost depends not on quantity but quality. Definitely a production machine requires more power than staging or ChatBot. It is not easy to say in advance how much computational power we need. Besides certain technologies can be bought for a lower price (e.g. RStudio Connect can cost about 2'000 EUR per YEAR as we are only a starting company).

Technology	Volume	Ceiling Total Cost
Domain name	1	20 EUR/year
Hosting	4-5	1000 EUR/month
DataBase	2 (30GB each)	3000 EUR/month
RStudio Connect	1 (R/Python APIs, RMarkdown, EDA)	1700 EUR/month
BB, Confluence, Jira	10 users	30 EUR/month

Table 5.3: Technology cost

## External Data Analysis

Before build any model we should evaluate the feasibility of our data product and raise a few important questions:

1. What are the most important factors to define the necessity of mental treatment? (it would be great to ask only relevant questions)
2. Is it possible to build any ML model based on those features to evaluate the risk of mental health problems? (as it is expected that all features are categorical)
3. What quality would be of a model if it is possible? (no signs that it is feasible to build a model with a decent balanced accuracy)

During search for external sources and methods of evaluation of mental health we found Open Source Mental Illness research (OSMI) to specify the questions for questionnaire and also use their data to evaluate the quality of a possible model.

The dataset (table 5.4) contains the following features (all are categorical except *age*; *treatment* was chosen as target variable):

## Data Preprocessing step

Briefly, Data preprocessing step includes such actions as

1. gender normalization (transformation to Male/Female/Other format)
2. removing age outliers (for certain entries value could be negative or more than 100)
3. checking missingness and removing features that are impossible to impute (only **comments** feature)
4. marking missing values and then imputing with knn approach where  $k = 3$

Feature name	Description
Timestamp	-
Age	-
Gender	-
Country	-
State	(only for US)
Self employed	Are you self-employed?
Family history	Family history of mental illness
Work interfere	Is mental health condition affects work?
No employees	The number of employees in your company or organization
Remote work	Having remote work (outside of an office) at least 50% of the time
Tech company	The employer is primarily a tech company/organization
Benefits	Providing mental health benefits by the employer
Care options:	Providing options for mental health care by the employer
Wellness program	Discussion about mental health as part of an employee wellness program by the employees
Seek help	Providing resources by the employer to learn more about mental health issues and how to seek help
Anonymity	Protecting anonymity if you choose to take advantage of mental health or substance abuse treatment resources
Leave	How easy is it for you to take medical leave for a mental health condition?
Mental-health consequence:	Having negative consequences caused by discussing a mental health issue with your employer
Phys-health consequence	Having negative consequences caused by discussing a physical health issue with your employer
Coworkers	Would you be willing to discuss a mental health issue with your coworkers?
Supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
Mental health interview:	Would you bring up a mental health issue with a potential employer in an interview?
Phys health interview	Would you bring up a physical health issue with a potential employer in an interview?
Mental vs Physical	Do you feel that your employer takes mental health as seriously as physical health?
Obs consequence	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
Comments	Any additional notes or comments
Treatment	Was there a treatment for a mental health condition?

*Table 5.4: Feature description table*

## Feature Selection step

Feature Selection step includes two steps: removing unimportant variables with Boruta algorithm and with correlation filtering. The result of Boruta can be found below in the picture.

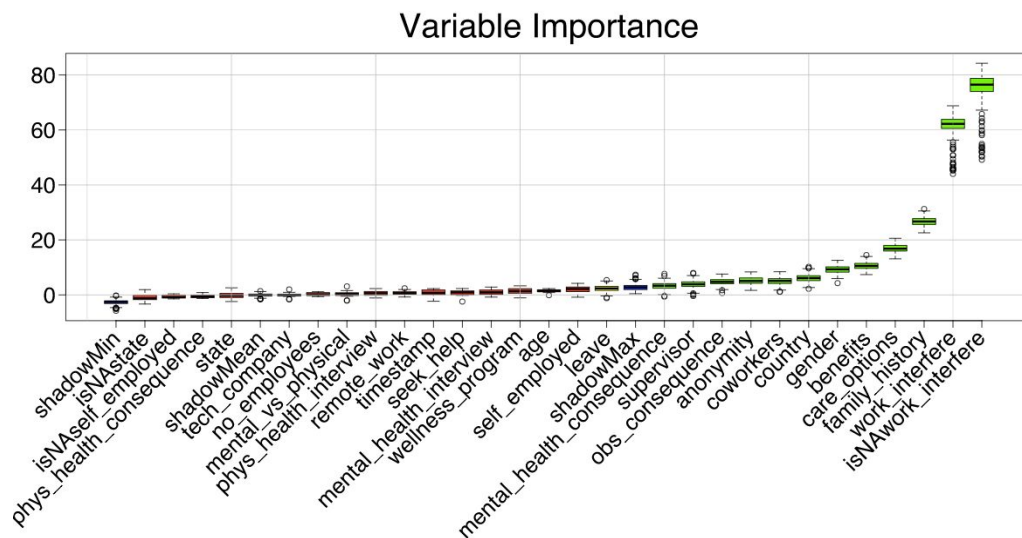


Figure 5.3: Boruta BoxPlot

In the figure 5.4 green whisker plots mark variables that should be taken into the model.

- Yellow variables are tentative variables.
- Red variables are unimportant.
- Blue variables are shadow metrics.

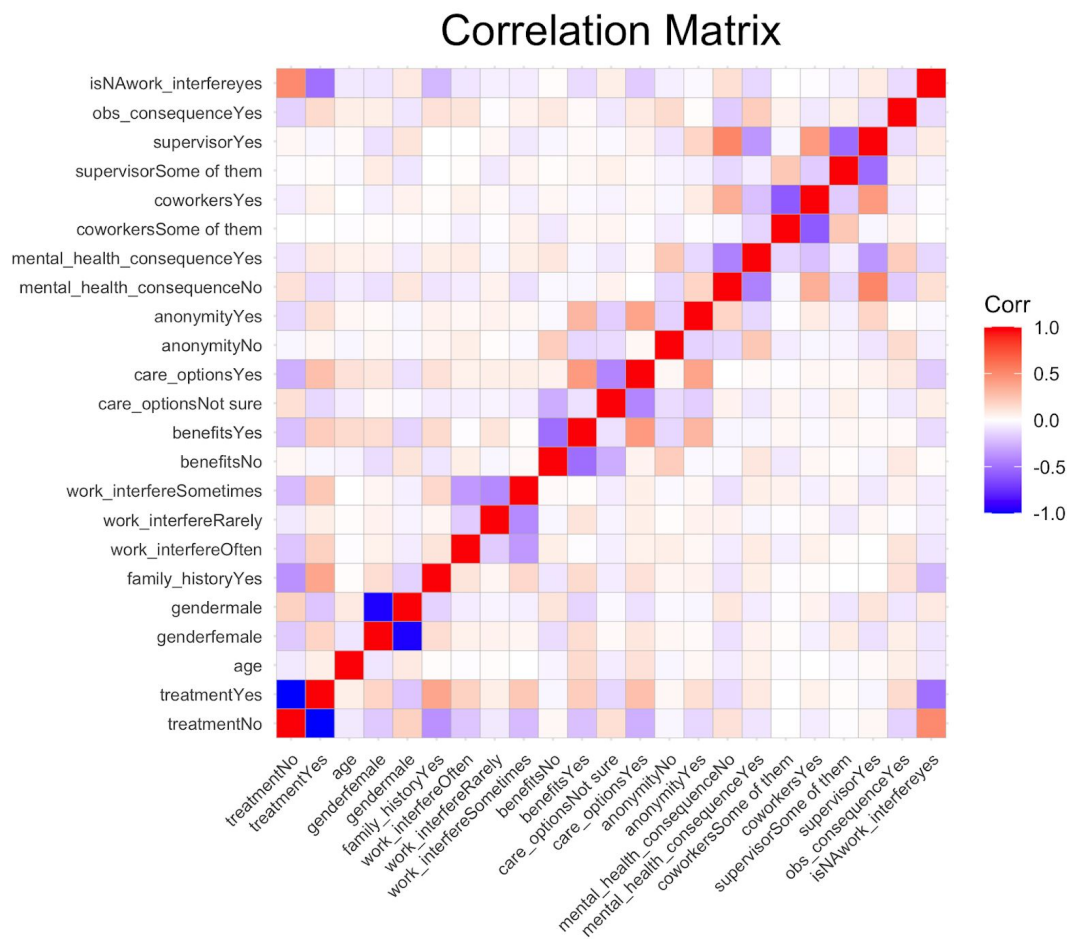
According to the plot above such variables as **work\_interfere** and **family\_history** are significantly more important than the rest. Also you can notice **isNAwork\_interfere** variable that shows what entries were imputed marked as the most significant. It is a good sign that confirms our initial hypothesis about strong dependence between mental state and its influence on the work.

After the Boruta feature selection step we look at correlation matrix to exclude other dependent variables.

In the figure 5.4 **country** feature and the rest features were separated to show these areas more closely. However there is no correlation between country and the rest so plotting those together would not make much sense.

You would notice a strong correlation between **treatmentNo** and **treatmentYes** as well as **genderfemale** and **gendermale**. Because these variables are complementary.

There is no need to exclude any variables by correlation value.



*Figure 5.4: Correlation Matrix*

The result of feature selection we have a narrowed dataset that has only the most important features:

- **treatment** (target)
- **country**
- **age**
- **gender**
- **family\_history**
- **work\_interfere**
- **benefits**
- **care\_options**
- **anonymity**
- **mental\_health\_consequence**
- **coworkers**
- **supervisor**
- **obs\_consequence**
- **isNAwork\_interfere**

## Modeling step: Hyperparameter Optimization

To predict necessity of treatment were considered three ML approaches:

1. Random Forest
2. Logistic Regression
3. Support Vector Machine

Unfortunately, a Logistic Regression based model is unable to handle that big amount of categorical variables that we have even if we use one hot encoding, so this model was excluded at an early stage. For a Random Forest Model, hyperparameter optimization was done regarding to such parameter as:

- **trees** - the number of trees contained in the ensemble.
- **min\_n** - the minimum number of data points in a node that are required for the node to be split further.

Whereas, the number of predictors that will be randomly sampled at each split when creating the tree models **mtry** was chosen as 3 as the dataset after feature selection step has only 14 features ( $m = \lfloor \sqrt{n} \rfloor$ , where  $n$  is the number of features and square brackets is the floor operator).

Random Forest penalty plot

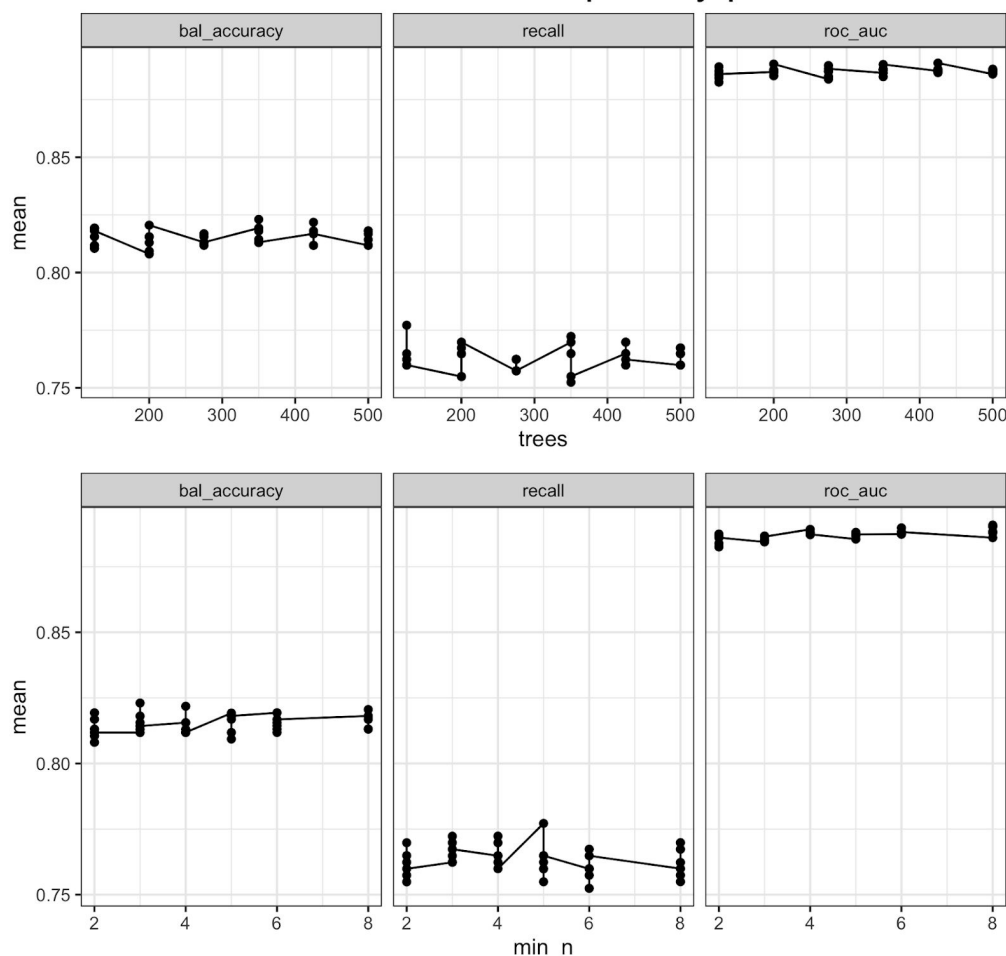


Figure 5.5: Random Forest Penalty Plot



The Penalty plot provides information about balanced accuracy, recall and ROC AUC. The parameters were chosen based on the last metric. The **trees** parameter equals 423 and **min\_n** is 8.

For a SVM model we used the radial basis function. Hyper parameter optimization was done regarding the following parameters:

- **cost** - the cost of predicting a sample within or on the wrong side of the margin.
- **rbf\_sigma** - the precision parameter for the radial basis function.

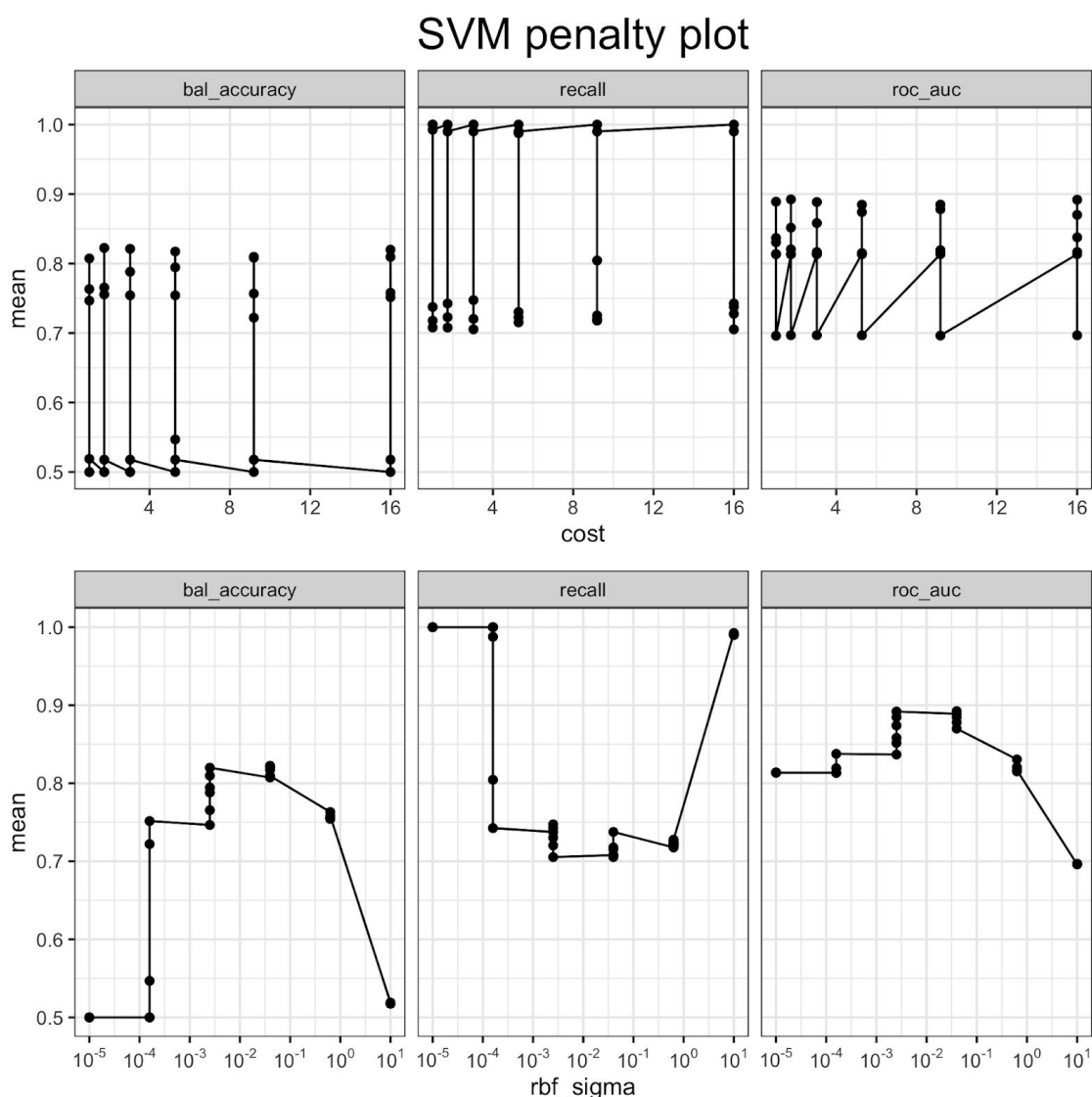


Figure 5.6: SVM Penalty Plot

Penalty plot provides information about balanced accuracy, recall and ROC AUC. The parameters chosen based on the last metric. The **cost** parameter equals approximately 1.7411 and **rbf\_sigma** is about 0.03981.

## Modeling step: Validation

For the Validation step to compare the results of two machine learning approaches we used ROC curves, distribution plots and such metrics as AUC ROC, Recall, Accuracy, Balanced Accuracy and Kappa. In the pictures below it can be seen that the difference between SVM and Random Forest is barely distinguishable. The AUC value is slighter higher for Random Forest.

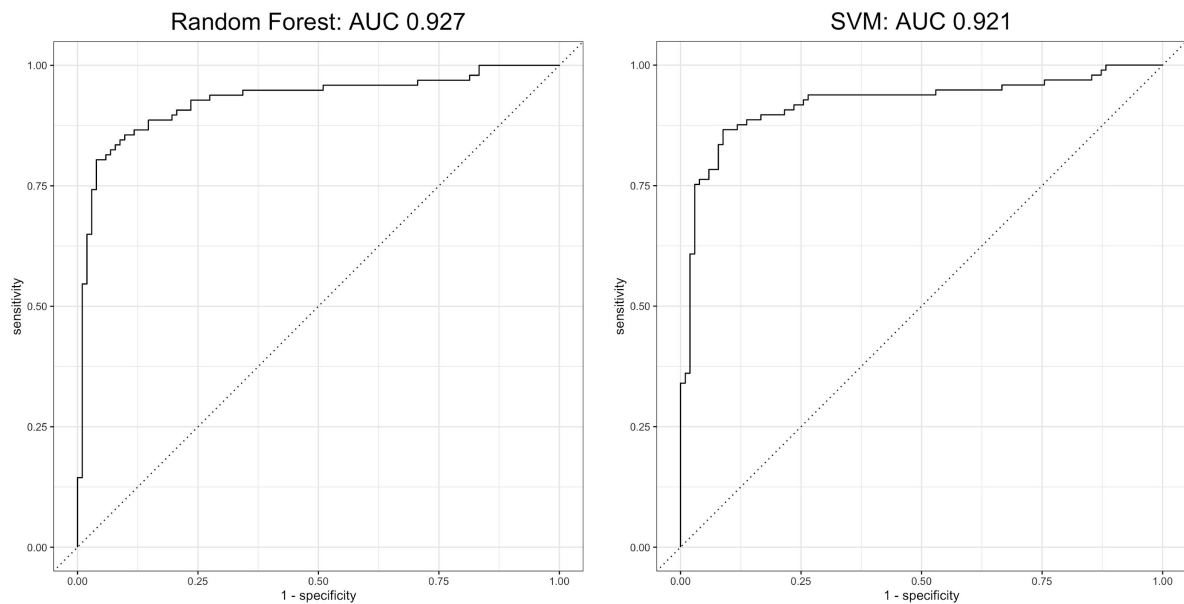


Figure 5.7: ROC comparison

Although comparing validation distributions both models show the appealing result as we can easily define a threshold at which classes are distinguishable, we can see that Random Forest classes are more skewed meanwhile SVM has a higher value of kurtosis.

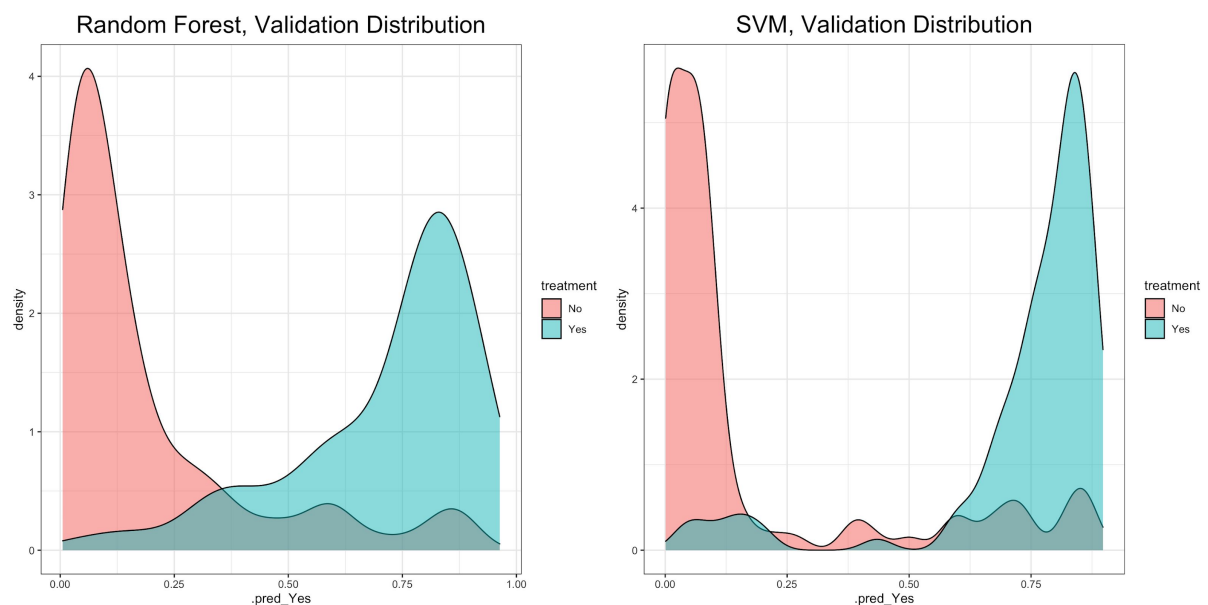


Figure 5.8: Distribution comparison

The table below allows us to compare two ML models with such metrics as AUC ROC, Recall, Accuracy, Balanced Accuracy and Kappa. All metrics say about moderate quality of the build models.

It is required an additional explanation for the recall metric which uses “No” value as a relevant and its high value for the SVM model confirms our earlier statement about kurtosis.

	<b>AUC ROC</b>	<b>Recall</b>	<b>Accuracy</b>	<b>Balanced Accuracy</b>	<b>Kappa</b>
<b>Random Forest</b>	0.927	0.853	0.869	0.870	0.739
<b>SVM</b>	0.921	0.922	0.864	0.863	0.728

*Table 5.5: Model Comparison Table*

The high value of Recall and high value of kurtosis tell us that the SVM model lacks generalization, so as one of the most possible consequences it can rather mark all unfamiliar data as one class rather than two different classes. Therefore it makes more sense to proceed with the Random Forest model and use this model as a baseline to use for comparison on different models.

## Modeling step: Testing

After taking the best hyperparameters of the Random Forest model and retraining it on the full training dataset and testing. We have the following results.

	<b>Recall</b>	<b>Accuracy</b>	<b>Balanced Accuracy</b>	<b>Kappa</b>
<b>Random Forest</b>	0.880	0.809	0.804	0.613

*Table 5.6: Final Model Comparison Table*

These results meet our expectation and it can be considered as a decent baseline for further research (see [https://github.com/ElijahOzhmegov/mental\\_health\\_analysis](https://github.com/ElijahOzhmegov/mental_health_analysis)).

## Conclusion

We used an external OSMI database to find the most important features and build two ML models, one of them could be used as a comparison baseline for further research. Besides that we did some data preparation, imputation and feature selection steps that lack explanation in the current research and some of them can be considered as controversial. Provided External Data Analys is strongly recommended to treat as an approximation of a research based on invalid data that was collected in the desired approach.

# Machine Learning Canvas










<b>Decisions</b>  How are predictions used to make decisions that provide the proposed value to the end-user?  <b>Based on our output the employer suppose to take measures to help an employee or in extreme cases, take a discharge actions.</b>  <b>Also it is expected that we will be aware what measures will be taken afterward via by HR or deployed psychologist.</b>	<b>ML task</b>  Input, output to predict, type of problem.  <b>Input: categorical and text based on questionnaire and performance review</b>  <b>Output: Is treatment required? (Yes/No)</b>  <b>How severely is an employee damaged? (no stress, acute stress, episodic acute stress, and chronic stress)</b> <b>Type: Classification</b>	<b>Value Propositions</b>  What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?  <b>1. Provide information if an employee experiences any mental issues which affect his/her performance</b>  <b>2. Increase employee productivity via by solving mental health issues before it is too late</b>	<b>Data Sources</b>  Which raw data sources can we use (internal and external)?  <b>1. internal: our data source is based on our questionnaire.</b>  <b>2. external: OSMI database</b>	<b>Collecting Data</b>  How do we get new data to learn from (inputs and outputs)?  <b>During Performance Evaluation, every 6 month from a particular company</b>
<b>Making Predictions</b>  When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?  <b>During Performance Assessment</b>  <b>Means of Internet Speed to provide a feedback (Instantly)</b>	<b>Offline Evaluation</b>  Methods and metrics to evaluate the system before deployment.  <b>Cross-Validation</b>  <b>ROC AUC, Recall, Balanced accuracy, Kappa</b>		<b>Features</b>  Input representations extracted from raw data sources.  <b>+ Categorized answers: Yes/No and 1-10</b>  <b>+ One text answer (possibly): it will require a sentimental analysis</b>  <b>+ Standard Performance Assessment text answers</b>	<b>Building Models</b>  When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?  <b>Every 3-6 months, depends on how many clients we have and when they do performance assessment, mostly it is expected during calm periods between assessments</b>  <b>It can take a few weeks and up to 6 months, strongly depends on the amount of data we are provided</b>
	<b>Live Evaluation and Monitoring</b> Methods and metrics to evaluate system after deployment, and to quantify value creation.	Dynamic assessment of an employee based on HR feedback:  If we take a certain employee and this person experiences mental health issues, we will notify about it an HR. Afterwards it is expected that the HR and the employee discuss it and will find a certain solution for that employee. HR will notify us about taken measures. And during the next assessment we will say determine if there is a better mental grade or not.		

Table 5.7: Machine Learning Canvas

## 6. Cash Flow Prediction

To predict our cash flow, we look at estimated cost (cash flowing out) and estimated revenues from our services (cash flowing in).

### Cost structure (Olena)

We have prepared a thorough cost budget for the first year of operation of our business. The link to the spreadsheet can be found here: [Cost Spreadsheet](#) (please click to request access).

All of the expenses are divided into sections, such as: Wages and Salaries; Office space costs; Insurance, fees and contributions; Taxes; Technical Expenses and others. Most of these costs are recurring and will remain as such every year. Other expenses such as company formation and registration fees - are one time expenses.

Most of the costs have been estimated, and this is important to disclose since we haven't started operating yet and don't know the actual costs. Some costs are very hard to estimate, for example taxes. We can only approximate our tax bracket, and that might change depending on our future profits or losses.

### Cash Inflow (Silke)

As we want to offer our product free of charge in the first year, we estimate revenues for the second year.

Internal financing is the result of a retained, positive cash flow. Our incoming cash flow would be generated from usage fees and email support services.

### Fees HappiEmployi

HappiEmployi will be offered to German medium sized companies for a yearly fee depending on the company's number of employees. Targeting medium sized companies we offer a pricing in tranches:

- 500 €/y for < 55 employees
- 1000 €/y for < 200 employees
- 1500 €/y for >= 200 employees.

Assuming our customers will be equally distributed over the tranches the expected revenue per customer / company is  $\frac{1}{3} * (500\text{€} + 1000\text{€} + 1500\text{€}) = 1000\text{€}$

Depending on our estimates of SOM as 1, 5 or 10% we could generate the following revenue

estimated SOM	Nr of customers / companies (acc. to x % SOM)	estimated revenue (basic service)	rounded estimated revenue
1%	940	940'000 €	1 Mil €
5%	4'700	4'700'000 €	5 Mil €
10%	9'400	9'400'000 €	10 Mil €

*Table 6.1: revenues basic service for estimated x% SOM*

## Fees email support

To ensure that HR personnel at our customers' companies use our service correctly and safely manage any technical challenges, we offer email support with different guaranteed response times at different prices. Our free email support comes with a 48-hours guaranteed response time. Shorter response times are offered:

Response time	yearly fee
10 hours	500 €
5 hours	1000 €
1 hour	2000 €

*Table 6.2: fees for guaranteed response times in email-support*

We estimate that half of our customers are satisfied with the 48h response time, and that the remaining half will have a tendency to the medium response time of 5 hours.

$$0.5 * (0.25 * 500€ + 0.5 * 1000€ + 0.25 * 2000€) = 0.5 * 1'125€ = 562.50€$$

estimated SOM	Nr of customers / companies (acc. to x % SOM)	estimated revenue (support)	rounded estimated revenue
1%	940	528'750 €	0.5 Mil €
5%	4'700	2'643'750 €	2.5 Mil €
10%	9'400	5'287'500 €	5 Mil €

*Table 6.3: revenues email support for estimated x% SOM*

The overall estimated revenue will then be (all values rounded in Mil €)

estimated SOM	estimated revenue (basic service)	estimated revenue (support)	overall est. revenue
1%	1	0.5	1.5
5%	5	2.5	7.5
10%	10	5	15

Table 6.4: overall revenues for estimated x% SOM, all values in Mil €

## Cash Flow (Silke)

With the estimated revenue stream and cost we can now estimate a cash flow for the first two years.

For the second year we see different growth scenarios in % of SOM driven by increased resources invested (=cost). We estimate that we can gain a 1% SOM with basically the same cost as in the first year. To achieve a higher % of SOM we have to invest more in marketing, technology and staff. We assume that cost would have to grow by factors of 2 (to gain 5% SOM) and 5 (to gain 10% SOM).

estimated SOM		total revenue (in Mil €)	cost (in Mil €)	cash flow (in Mil €)
	year 1	0	1	- 1
1%	year 2	1.5	1	0.5
5%	year 2	7.5	2	5.5
10%	year 2	15	5	10

Table 6.5: cash flow for different growth scenarios

So with an estimated SOM of 5% for the second year our company will make a positive cash flow of about 5.5 Mil €, immediately setting off the negative cash flow in the first year of about 1 Mil €.

## Investment Capital (Olena)

Our estimated costs for the first year of doing business are € 965,000. We are looking for an investment capital to cover at least the first two years of operating:

2 Years x Annual Costs = € 2M

Therefore we are asking our investors for € 2 million of an initial series A round of funding. This would guarantee us an extensive development of our software product while establishing our brand on the market.

## 7. Legal structure (Silke)

### UG & GmbH

We, the team of HappiEmployi, have recently founded the HappiEmployi UG (haftungsbeschränkt), a German company of limited liability. Since we are all poor students, we started with 1€ as the company's capital. This contribution was made equally by all team members and the company is now shared equally between the partners.

As the company generates revenues we will yearly retain 25% of revenues until 25,000€ are accumulated. We will then transform our company into a GmbH.



# References

Teo

<https://next.canvanizer.com/demo/wA-82vhUOBc>

<https://www.w3schools.com/>

Yusif

<https://startuptalky.com/cac-by-industry/#:~:text=I'm%20sure%20you%20know,total%20company%20revenue%20on%20marketing>.

[Business Model Navigator](#)

[The Ansoff Model](#)

Olena

<https://www.marketdataforecast.com/market-reports/mental-health-software-market>

<https://www.lightercapital.com/blog/what-is-total-addressable-market-tam/>

<https://www.eu-startups.com/2019/06/10-european-startups-revolutionizing-mental-health/>

<https://www.statista.com/statistics/462234/it-industry-number-of-companies-germany/>

[https://www.g2.com/categories/mental-health?utf8=%E2%9C%93&selected\\_view=grid&segment=all#grid](https://www.g2.com/categories/mental-health?utf8=%E2%9C%93&selected_view=grid&segment=all#grid)

Ilia

<https://osmihelp.org/research>

<https://www.machinelearningplus.com/machine-learning/feature-selection/>

<https://medium.com/louis-dorard/from-data-to-ai-with-the-machine-learning-canvas-part-iii-868fe17b9be6>

<https://www.digitalocean.com/pricing/>

<https://aws.amazon.com/redshift/pricing/>

<https://aws.amazon.com/rds/postgresql/pricing/?pg=pr&loc=3>

<https://www.caissarecruitment.com/blog/berlin-tech-salaries-2019-how-much-do-developerssoftware-engineers-make-berlin/>