WELCOME!

Sebastian Siegler

Economist
Work experience in
logistics sector

Dr. Karima Chakroun

Cognitive Neuroscientist
Psychologist
Biochemist

Today: Consulting Data Scientist at FutureMinds

Health Insurance Premium

Calculations based on psychological predictors for drug abuse risk

TABLE OF CONTENTS

01 Problem Statement

02 Business Value

03 Methodology

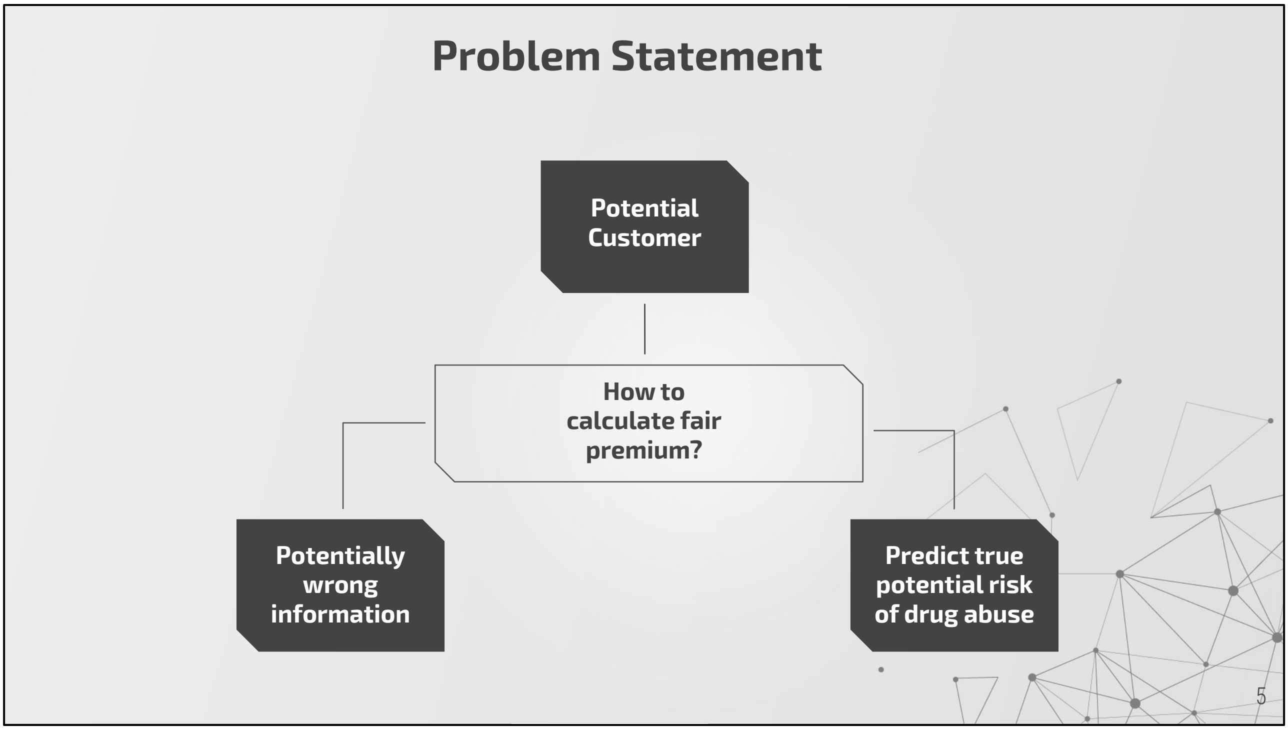
04 Conclusion

05 Future Work

A Appendix

01

Problem Statement



Problem Statement

Potential Customer

How to calculate fair premium?

Potentially wrong information

Predict true potential risk of drug abuse

Health insurance companies have to calculate a premium for their products that make them competitive at the market while ensuring appropriate profits.

Drug abuse causes immense costs for the healthcare sector.

Potential customers are dishonest when it comes to admitting previous drug use.

It is of great interest for us to predict a potential drug abuse risk for potential customers of private health insurances in order to calculate a fair premium.

02

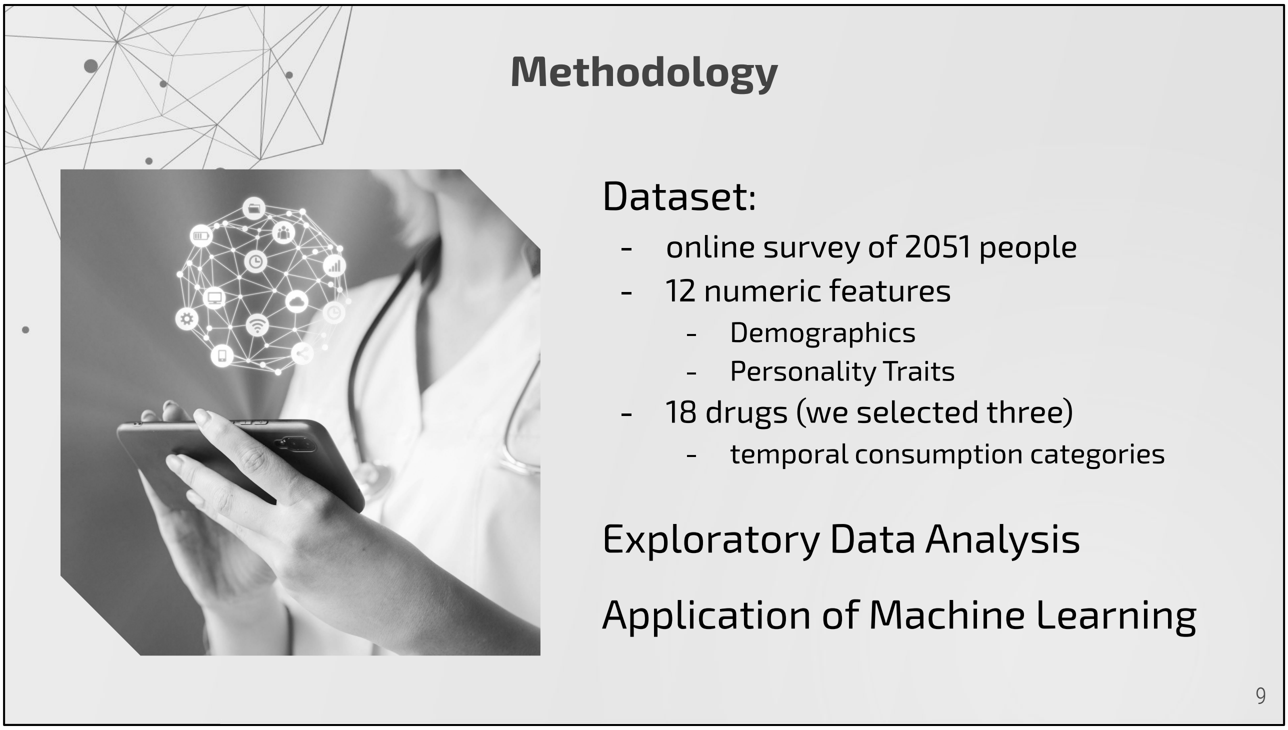
Business Value

Predict potential costs caused by customers with drug abuse problems (treatment, sick leaves, …) in order to:

- reject customers who potentially consume heavy drugs like cocaine
- charge customers who potentially consume moderately heavy drugs like speed or cannabis a risk premium
- accept customers who do not consume any drugs with the basic insurance premium

03

Methodology



Methodology

Dataset:

- online survey of 2051 people
- 12 numeric features
  - Demographics
  - Personality Traits
- 18 drugs (we selected three)
  - temporal consumption categories

Exploratory Data Analysis

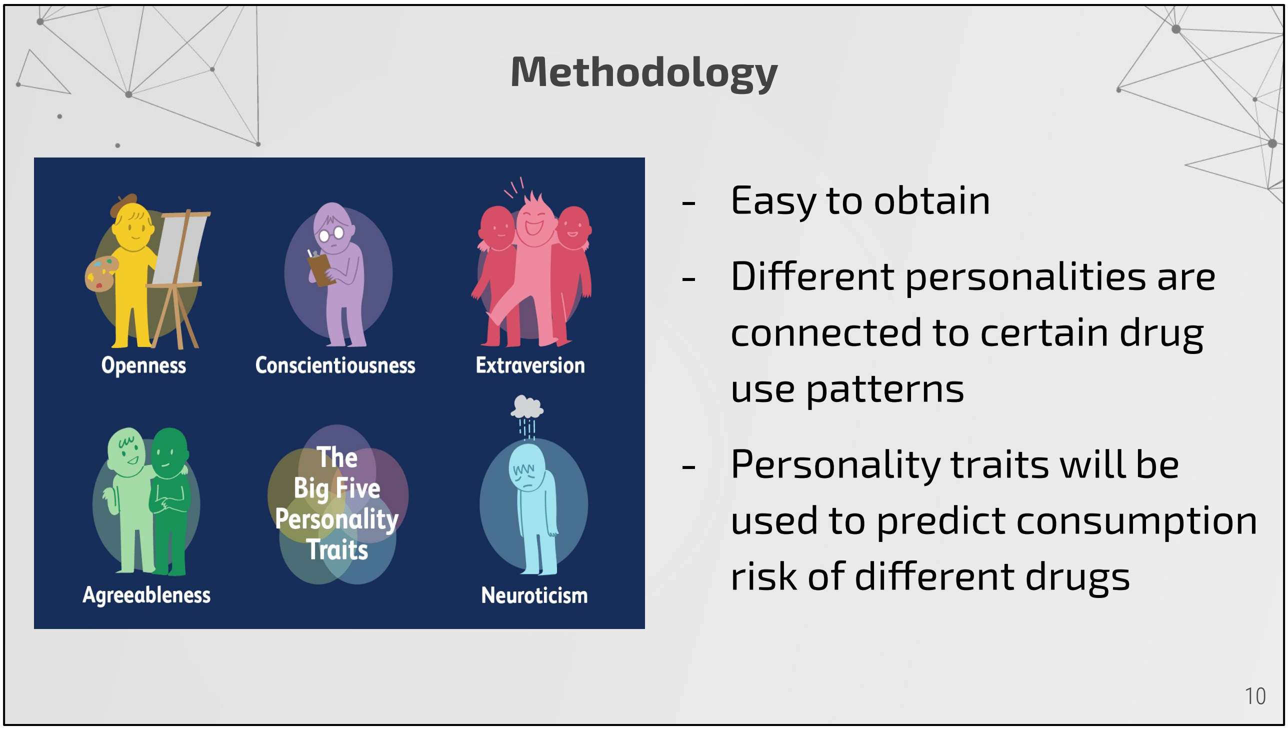
Application of Machine Learning

Data source:

- we were able to use data collected by research facilities
  - 'The Five Factor Model of personality and evaluation of drug consumption risk' (Fehrman et al., 2017)
  - link to research article: https://arxiv.org/abs/1506.06297
  - link to dataset and data dictionary: https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36536
- anonymous online survey, ran over 12 months, snowball sampling method
- 2051 took part in the survey, 166 excluded due to incorrect answers or inattentiveness

Dataset:

- 12 features: all numeric based on numerical/categorical feature quantification as described in research article
  - 5 demographic variables: age, gender, education, country, ethnicity
  - 7 personality traits:
    - Big Five: Neuroticism, Extraversion, Openness to Experience, Agreeableness, Conscientiousness
    - Impulsivity, Sensation Seeking
- 18 drugs: categorized into 7 different temporal consumption categories:
  - **CL0:** Never Used, **CL1:** Used over a Decade Ago, **CL2:** Used in Last Decade
  - **CL3:** Used in Last Year, **CL4:** Used in Last Month, **CL5:** Used in Last Week, **CL6:** Used in Last Day



Methodology

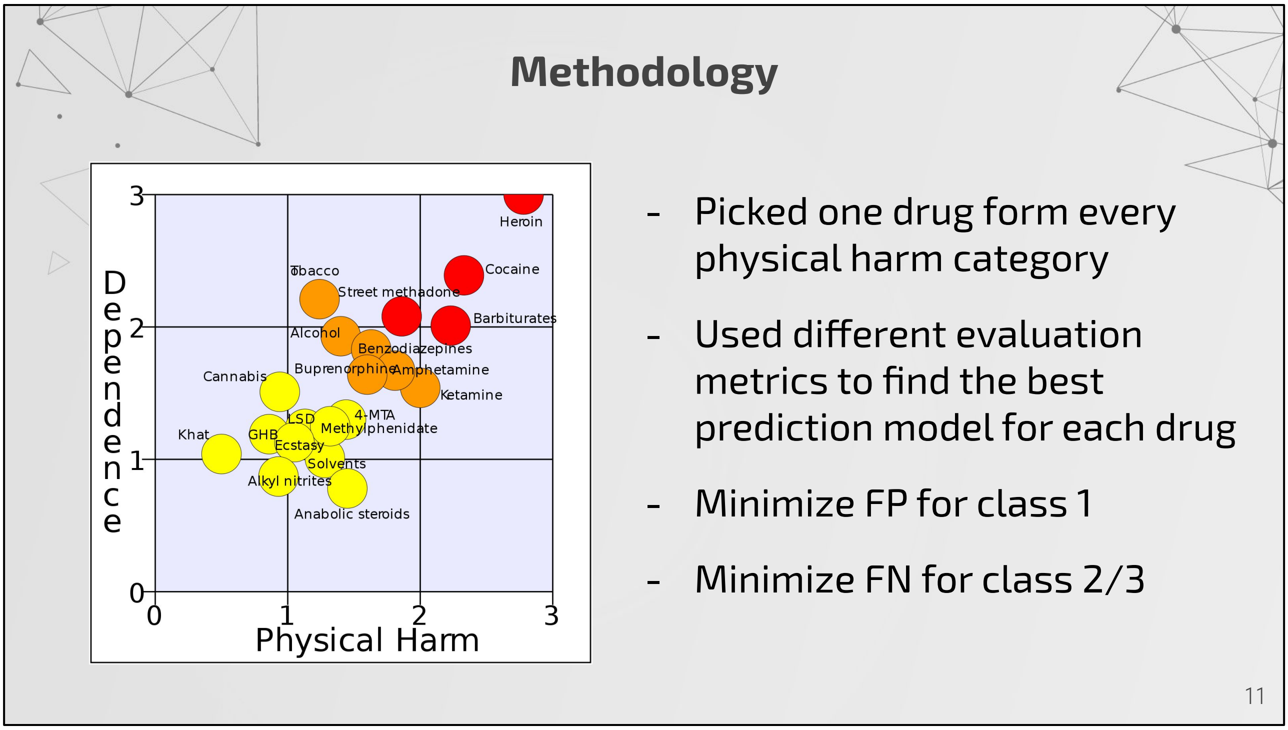
- Easy to obtain
- Different personalities are connected to certain drug use patterns
- Personality traits will be used to predict consumption risk of different drugs

The ‘Big Five’ personality traits are easy to obtain via questionnaire.
These personality traits were shown to be connected to certain drug use patterns (see research article by Fehrman et al., 2017) and will be used to predict consumption risk of different drugs.

Potential questions to measure ‘Big Five’:

- **Openness**: “I have excellent ideas.”, “I am quick to understand things.”
- **Conscientiousness**: “I am always prepared.”, “I pay attention to details.”
- **Extraversion**: “I don't mind being the center of attention.”, “I feel comfortable around people.”
- **Agreeableness**: “I am interested in people.”, “I sympathize with others' feelings.”
- **Neuroticism**: “I am relaxed most of the time.” (*reversed*), “I seldom feel blue.” (*reversed)*

Image source: https://www.verywellmind.com/the-big-five-personality-dimensions-2795422



Methodology

- Picked one drug form every physical harm category
- Used different evaluation metrics to find the best prediction model for each drug
- Minimize FP for class 1
- Minimize FN for class 2/3

For our business case, we have different requirements concerning the evaluation metrics of our models, depending on the physical damage potential of the drug.

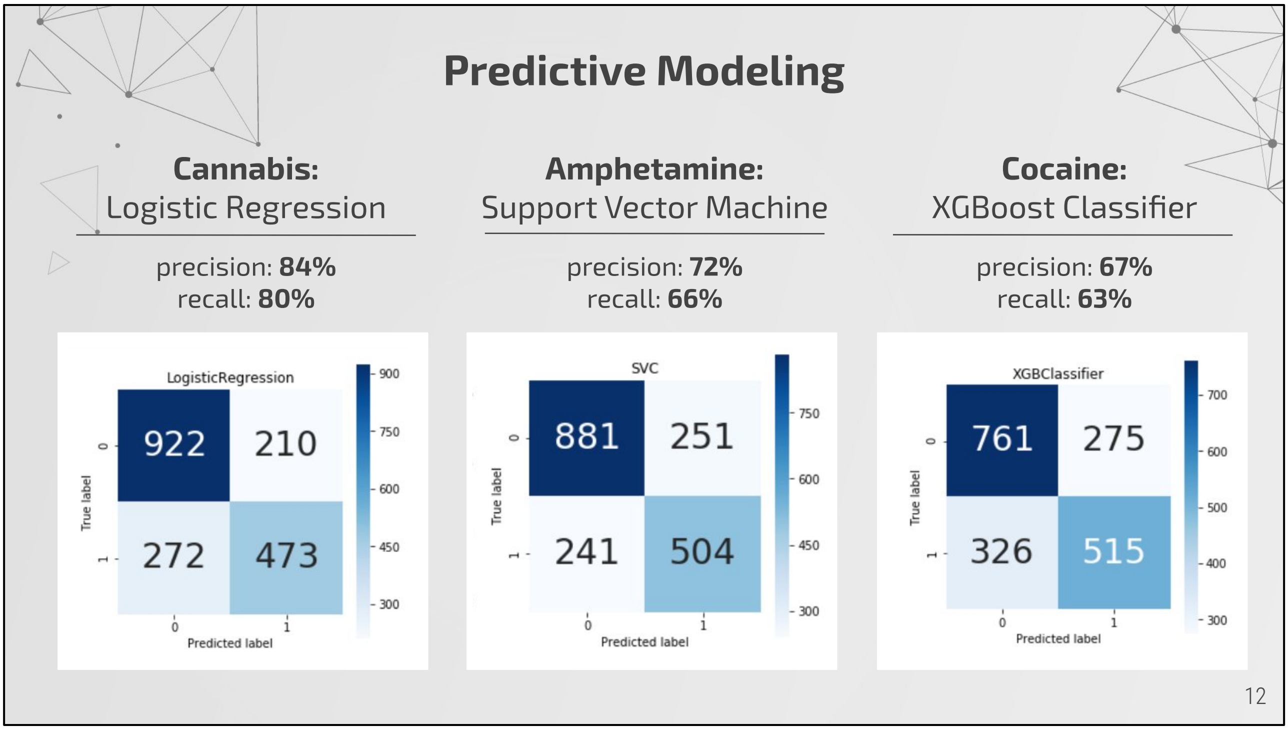
1) moderate (amphetamine) and very high (heroin)

- we want to correctly identify customers with a high probability of taking very harmful drugs due to the high expected treatment costs associated with that drug usage
  - in case of cocaine, we want to completely reject the customer because the insurance fees will not cover potential treatment costs
  - in case of amphetamine, we want to charge the customer an additional risk premium in order to minimize the potential loss due to costs associated with drug abuse
- in both cases we therefore want to avoid classifying a person as non-user, who has actually already consumed the drug or has a high probability to do so in the future
- we want to **minimize false negatives (FN)**, hence we want to **maximize recall**

2) low (cannabis)

- in this case, the risk premium we charge will cover the potential treatment costs associated with the consumption of the drug, hence we are indifferent when it comes to accepting these customers
- but we want to maximize our client pool and profits, therefore we don't want to charge an additional premium to potential customers falling into this category because this might result in losing the customer to the competitors who classifies the customer correctly and doesn't charge a risk premium
- we therefore want to avoid classifying a person as user, who has actually never consumed the drug or has no risk of doing so in the future
- we want to **minimize false positives (FP)**, hence we want to **maximize precision**

Image source: https://www.nateliason.com/blog/drugs



Predictive Modeling

Cannabis:
Logistic Regression

precision: **84%**
recall: **80%**

Amphetamine:
Support Vector Machine

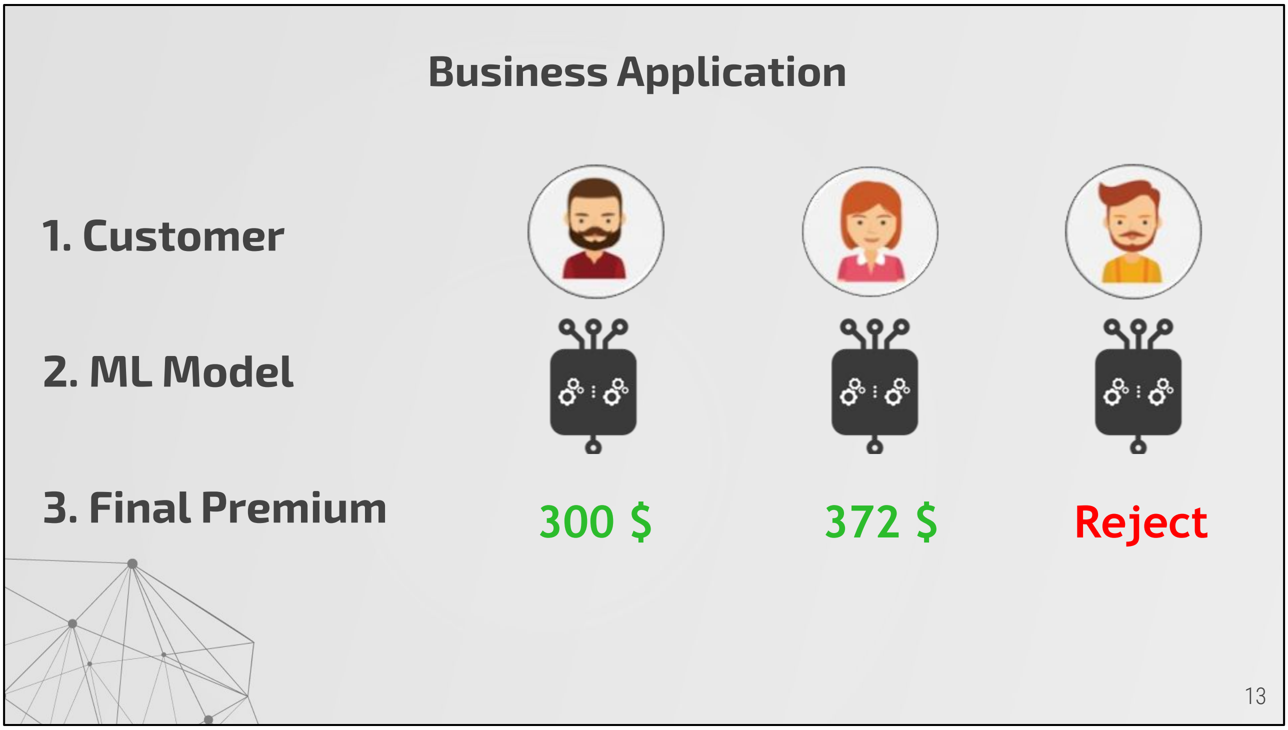
precision: **72%**
recall: **66%**

Cocaine:
XGBoost Classifier

precision: **67%**
recall: **63%**

- based on demographic and personality data, we are able predict the probability of potential drug abuse for different drugs with good precision and recall
  - different evaluation metrics (i.e. precision vs. recall) are important for our business case (see above)
  - different classifiers yield different results in terms of evaluation metrics

- based on the drug-specific requirements, different classifiers were selected for each drug:
  - **Logistic Regression** for Cannabis: yielding **84% precision** and **80% recall**
  - **Support Vector Classifier** for Amphetamine: yielding **72% precision** and **66% recall**
  - **XGBoost Classifier** for Cocaine: yielding **67% precision** and **63% recall**



Business Application

1. Customer
2. ML Model
3. Final Premium

300 $

372 $

Reject

In our business case, we can now use our winner models for each drug to make predictions of potential drug abuse in order to:

- calculate a risk premium added to the basic fee based on drug consumption risks
- completely reject a customer (in case of predicted heavy drug abuse)

The business application includes the following steps:

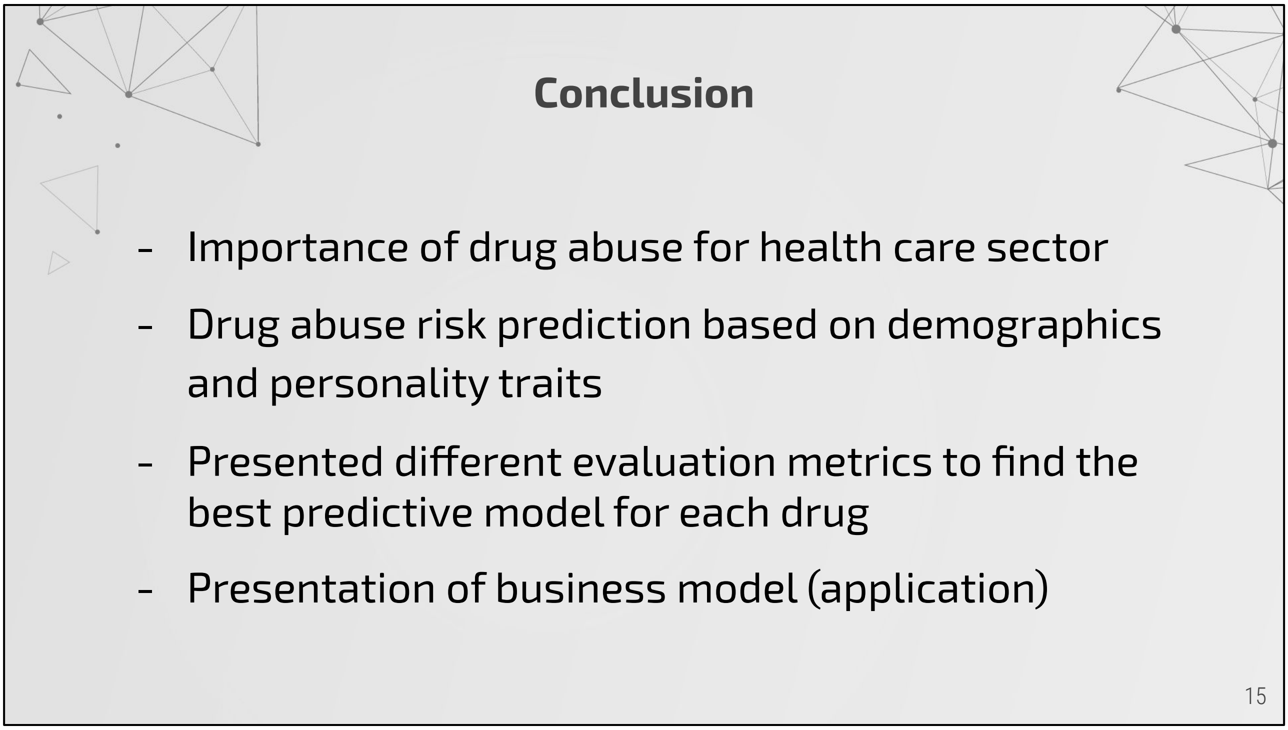
1. Collect demographic data and personality trait scores of a potential customer via (online) questionnaires
2. Using these inputs, our classifiers will make predictions on the drug consumption risks of this customer for different drugs (Cannabis, Amphetamine, Cocaine - only 3 drugs for now, as a proof of principle)
3. It either returns the final health insurance premium or - in the case of a high (>50%) consumption risk for heavy drugs - that the customer needs to be rejected

Image sources:

- https://www.freepik.com/free-vector/colorful-collection-with-great-variety-avatars_1258263.htm#page=1&query=avatar&position=0
- https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

04

Conclusion

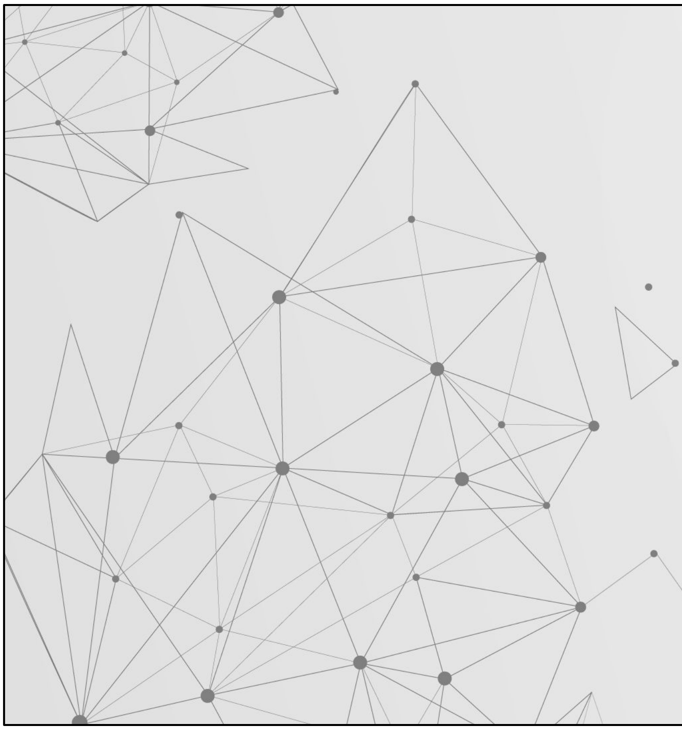


Conclusion

- Importance of drug abuse for health care sector
- Drug abuse risk prediction based on demographics and personality traits
- Presented different evaluation metrics to find the best predictive model for each drug
- Presentation of business model (application)

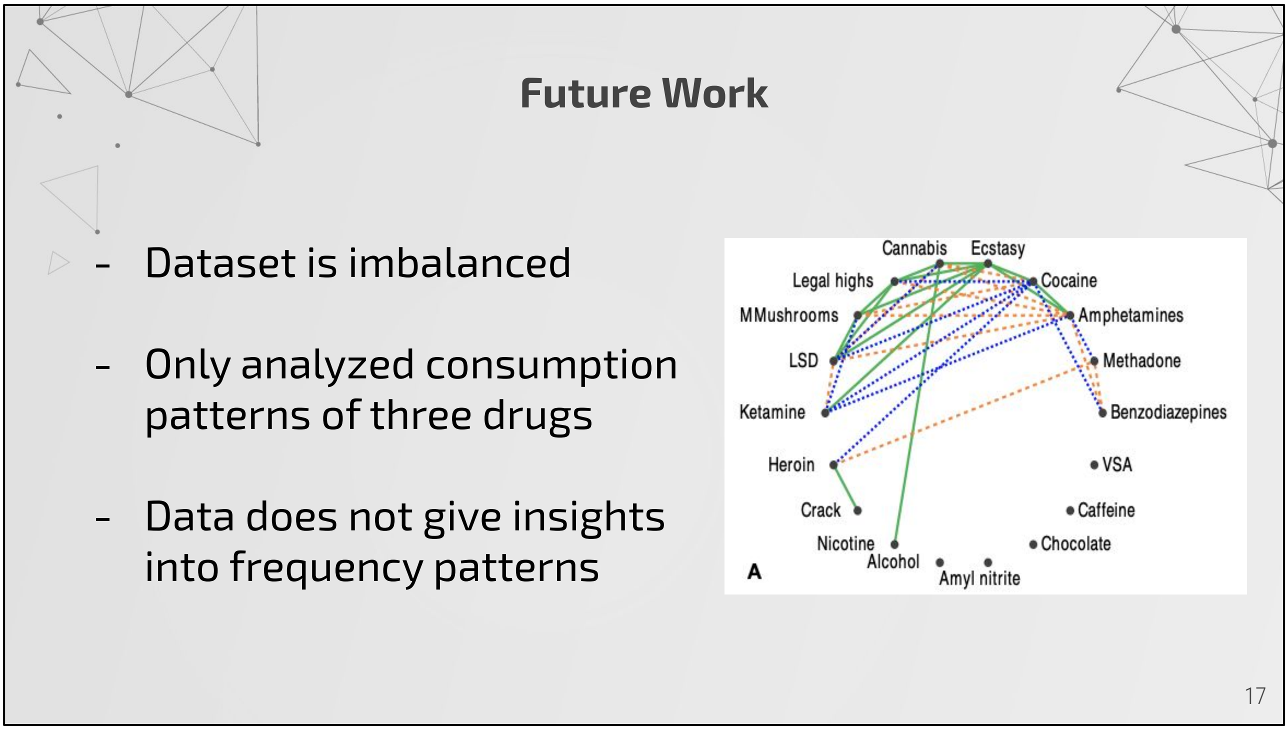
Summary and conclusions:

- Importance of drug abuse for health care sector
- Drug abuse risk prediction is possible based on questionnaire data (demographics and personality traits)
- We presented different evaluation metrics to find the best predictive model for each drug
- We presented a business application based on the modeling results, which calculates health insurance premiums based on drug consumption risks



05

Future Work



Future Work

- Dataset is imbalanced
- Only analyzed consumption patterns of three drugs
- Data does not give insights into frequency patterns

Cannabis
Ecstasy
Legal highs
Cocaine
MMushrooms
Amphetamines
LSD
Methadone
Ketamine
Benzodiazepines
Heroin
VSA
Crack
Caffeine
Nicotine
Chocolate
Alcohol
Amyl nitrite
A

Some prospects for future work include:

- deal with imbalanced dataset (for many drugs much more non-users than users)
  - use drug groups/pleiades based on correlation between drug consumption
  - might yield more robust predictions than using individual drugs for predictions
- collect more data
  - especially about the **frequency** of drug use, not only the last time point of drug use
  - about the relapse risk for different drugs and the associated economic costs
- fine-tuning of model (hyper)parameters via grid search:
  - different k and distance metrics for KNN
  - different maximum depths for classifiers based on decision trees
  - different kernels for SVC (linear/non-linear)

Image source: https://arxiv.org/pdf/1506.06297.pdf

THANKS

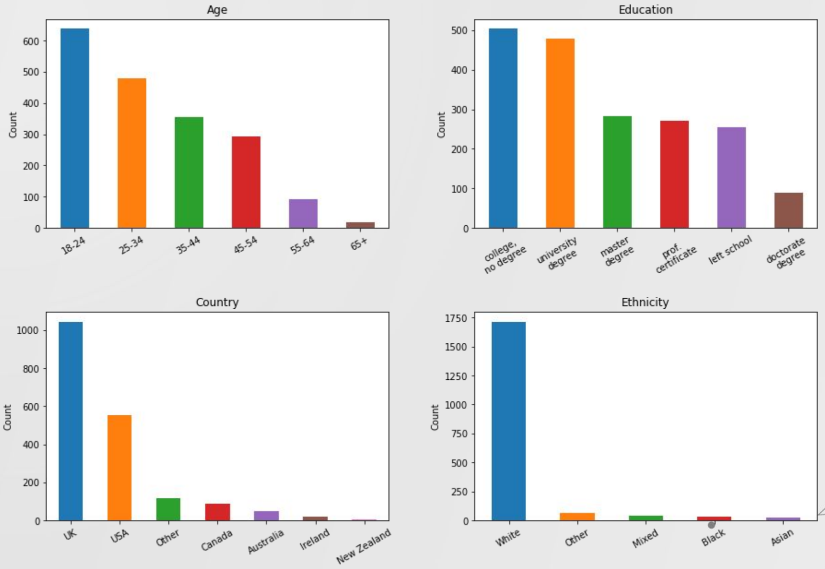
CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.

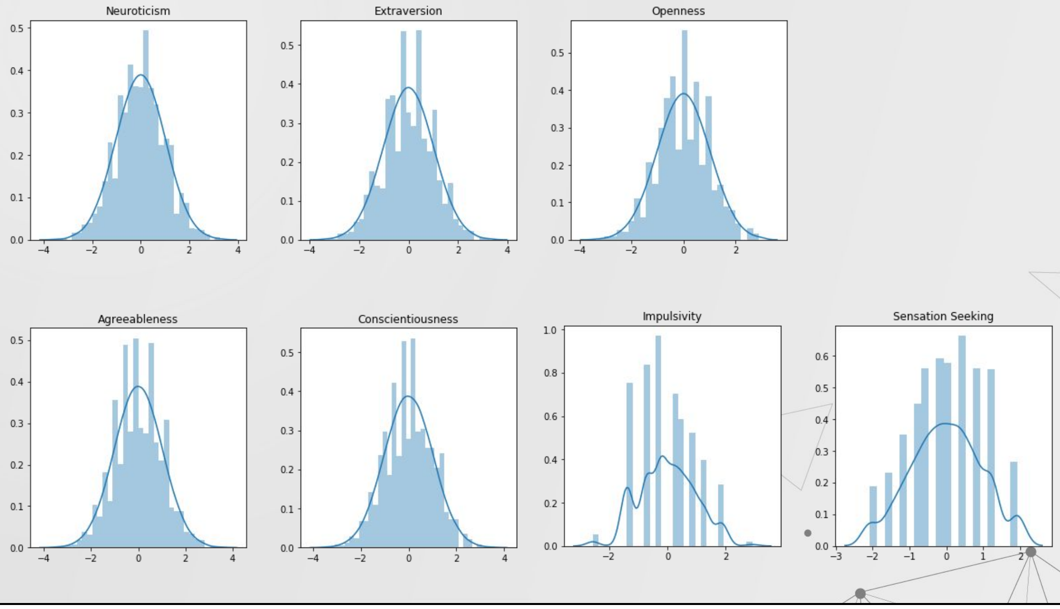
A

Appendix

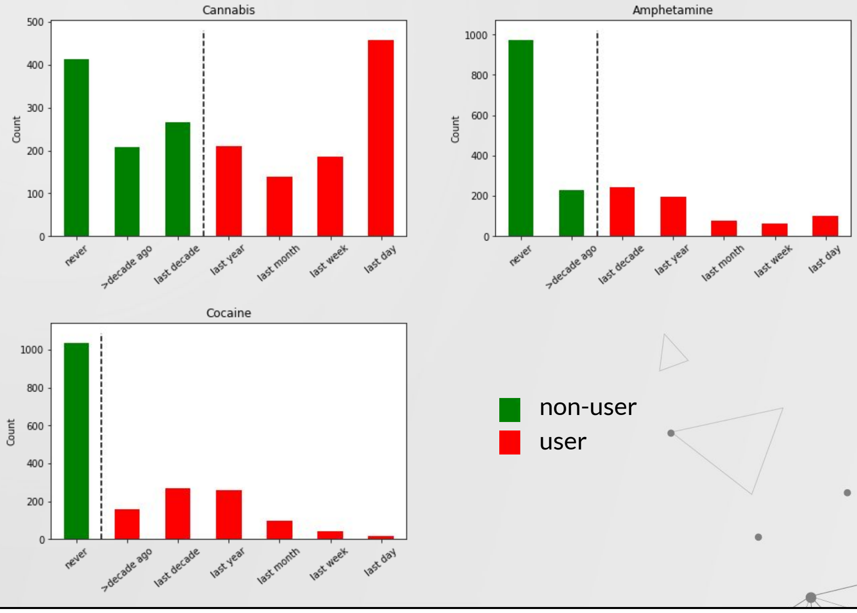
Demographics: distribution



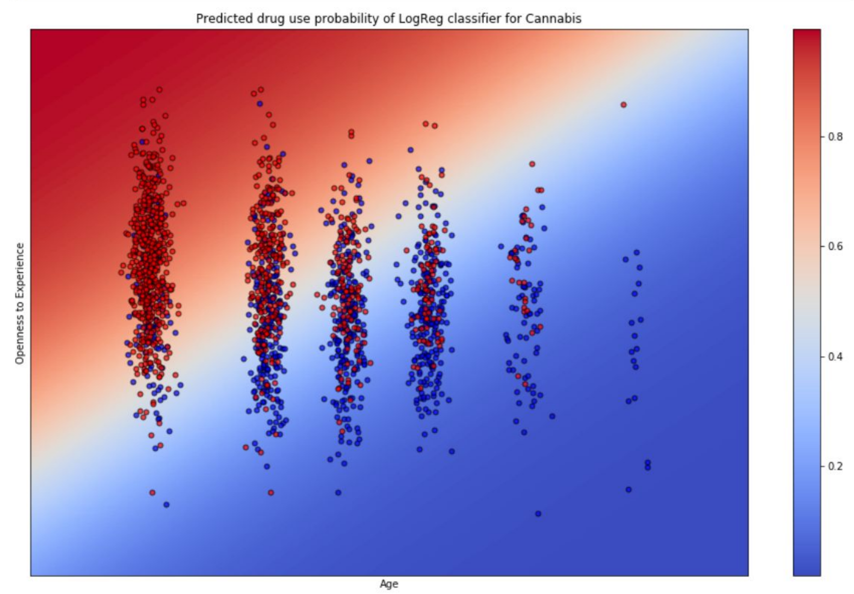
Personality traits: distribution



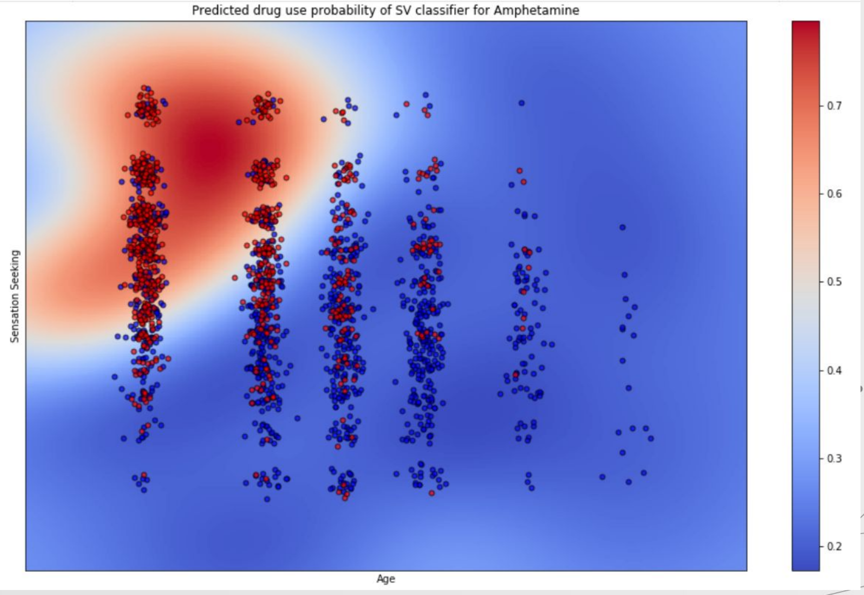
Drug consumption patterns



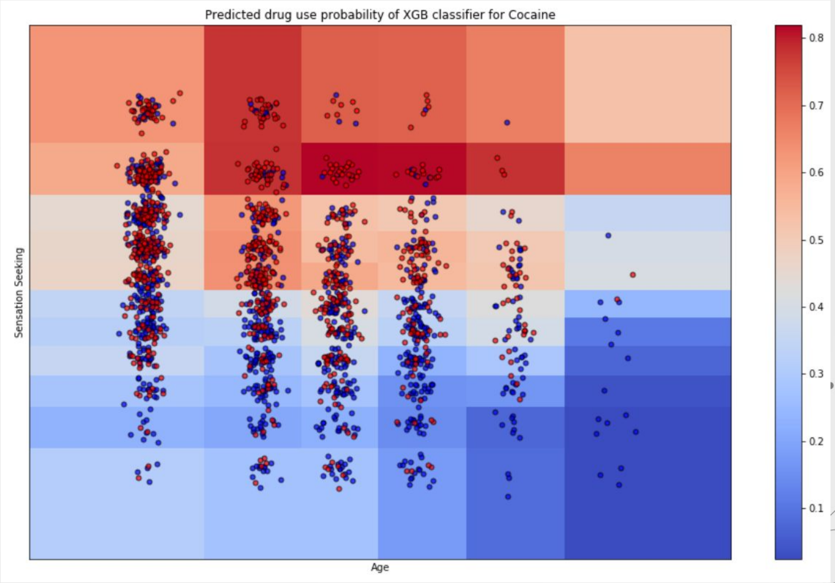
Cannabis use: Logistic Regression Model



Amphetamine use: Support Vector Classifier



Cocaine use: XGBoost Classifier


Predicted drug use probability of XGB classifier for Cocaine
Sensation Seeking
Age