**Prudential Financial**

**Data Scientist Recruitment Exercise**

Thank you for applying for this position at Prudential Financial and congratulations on getting through to

this stage of our interview process!

By joining Prudential, you become part of a growing, dynamic company that values and supports its

talent and not to mention, great compensation, work/life balance, paid time off and generous benefits.

We’re proud to be part of a company that supports its talent through challenging work and a focus on

development.

In preparation for your on‐site interview, we ask you to demonstrate your data science skills on an analytics exercise.

**Why an analytics exercise?**

We’re interested in your ability to solve real‐world data science problems, not in how well you perform

in the artificial interview environment. Therefore, at this stage of the interview process,

we’d like to simulate a real‐world problem similar to something you might encounter when working at

Prudential.

As a side benefit, you can work on the problem using familiar tools and environment at a time that suits you. We’ve selected an exercise that should take you about 3‐4 hours to complete.

**Questions**

You shouldn’t have any questions about the task. If you find yourself unsure about something; then

please make an appropriate assumption and explain/document your rationale. Please note that these

instructions are intentionally vague so as to better simulate real life data science problems and your skill

in navigating them…

**Instructions**

Assume you are an employee at the hypothetical company “GoodHealthCo” and you have received the

email included in the following section titled “**The Email from Dr. Snow**”.

Your task is reply to Dr. Snow’s email. Rather than email your response to Dr. Snow – **please email your**

**response to your Prudential HR representative!!!**

**Onsite Task**

During your onsite interview onsite at Prudential you’ll be asked to give a presentation

of your response to our recruitment team: you should imagine this as an informal presentation to a mix

of GoodHealthCo SBU leaders and your GoodHealthCo data science colleagues.

The format will be roughly 30 min presentation then 30 min questions & discussion – we’ll be sitting around a table and have an electronic monitor available if you’d like to use it.

**Copyright – do not share!**

This material is copyright © Prudential Financial. To keep this interview process fair for all applicants we

require that you don’t distribute, discuss, email, post, blog or otherwise share this exercise, the data set,

your response or any questions or discussion we have during telephone or onsite interviews.

© Prudential Financial

**The Email from Dr. Snow**

**To: analytics\_candidate@goodhealthco.com**

**From: Dr.John.Snow@goodhealthco.com**

**Subject: Fwd: Predicting Health Status Misrepresentation**

Hi candidate,

I’ve just received the email below from Monty Hall, the head of the Survey Business Unit (SBU).

The SBU is seeking our assistance with a project to detect misrepresentation in a survey data set. They

have emailed me a number of questions and an example data set.

Could you please spend 3‐4 hours (only!) looking at their questions and data and prepare an email

response.

Your task is not to completely solve the problem, but rather to provide a preliminary opinion on the task

and how we in the analytics team could help the SBU solve their problem.

Our audience in the SBU don’t understand code (R, SAS, Python, etc) so your response should be a short,

clear and concise email of minimal length.

If you’d like to provide code or detailed results for the benefit of your analytics team colleagues, then

you should append it as a separate document.

Thanks,

Dr John Snow,

Head of Analytics

GoodHealthCo

**FORWARDED EMAIL‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐‐**

**To: analytics\_team@goodhealthco.com**

**From: monty\_hall@SBU.goodhealthco.com**

**Subject: Predicting Health Status Misrepresentation**

Hi Dr Snow,

The GoodHealthCo Survey Business Unit (SBU) runs an annual telephone survey called the “Behavioral

Risk Factor Surveillance System”. There’s further information about the survey in the package

accompanying this document.

© Prudential Financial

**The SBU is concerned that some survey respondents may be intentionally misrepresenting their**

**health status for Diabetes “DIABETE3” (variable number VARNUM 45) and therefore reducing the**

**value of the survey.**

The SBU believes that we may be able to identify misrepresentation by predicting an individual’s health

status from other available variables; the idea being that honest responses will be consistent with other

information we have about the respondent.

Thus significant differences between predicted status and actual survey response may indicate risk of

misrepresentation that could then be used to target some other mechanism for validating the accuracy

of their response (such as a medical test or medical records).

We had an intern develop a preliminary model which we’ve included FYI at the end of this email.

I’m hoping you may help us understand this problem and would appreciate your advice and next steps.

Also, we have a few specific questions which we’ve included below.

Thanks,

Monty Hall,

Head of GoodHealthCo SBU

© Prudential Financial

Q1. Intern analysis

The SBU had an intern look at the predictive problem; the intern’s model output is included at the end

of this email. The model seems to be very good, can you please comment on its suitability for our task?

Q2. Predictive Model

The SBU would like some indication as to how accurately the analytics team believe they can predict

DIABETE3 when using \*only\* the data acquired from the telephone survey. (i.e. only fields in the

accompanying data file).

*In this question we’re looking for your ability to \*quickly\* produce an approximate model and*

*communicate it while fully cognizant of the model’s limitations and risks; this carries more weight in our*

*considerations than the actual model performance you achieve relative to the performance achieved by*

*other candidates. Use analytical tools of your choice.*

Q3. Improving predictive model

If the SBU decided to go ahead with this project, then concisely and in priority order list what model

improvements should be explored.

*You may assume there are no data or technology constraints at GoodHealthCo; however, you should be*

*conscious of the value created by costs that are additional to the SBU’s current cost of conducting the*

*telephone survey.*

© Prudential Financial

Q4. Identifying misrepresentation

What are your thoughts about the SBU’s misrepresentation concerns and their plans for detecting

misrepresentation? How would you go about solving the SBU’s misrepresentation concerns?

Q5. Managing the project

If the SBU was to work with the analytics team; how should the project be managed?

Q6. Are there any questions you would like to ask the SBU?

Do you have any clarifying questions you would like to ask the SBU? Please list your questions in order of

importance and explain why the question is important.

The Intern’s analysis:

**set.seed**(917);

Data <‐ Data\_Model[**sample**(**nrow**(Data\_Model)),]

train <‐ Data[1:**floor**(0.7\***nrow**(Data)),]

test <‐ Data[(**floor**(0.7\***nrow**(Data))+1):**nrow**(Data),]

**library**(caret)

**library**(randomForest)

## randomForest 4.6‐10

## Type rfNews() to see new features/changes/bug fixes.

##

## Attaching package: 'randomForest'

##

## The following object is masked from 'package:ggplot2':

##

## margin

rf\_model <‐ **train**(**factor**(Diabetic\_Ind) ~ ., data=train, method="rf")

rf\_model

## Random Forest

##

## 4805 samples

## 13 predictor

## 2 classes: '0', '1'

##

© Prudential Financial

## No pre‐processing

## Resampling: Bootstrapped (25 reps)

## Summary of sample sizes: 4805, 4805, 4805, 4805, 4805, 4805, ...

## Resampling results across tuning parameters:

##

## mtry Accuracy Kappa Accuracy SD Kappa SD

© Prudential Financial

## 2 0.9298018 0.5663685 0.0056124887 0.029080037

## 35 0.9992963 0.9966816 0.0005503165 0.002649026

## 68 0.9989349 0.9949897 0.0007756410 0.003697180

##

## Accuracy was used to select the optimal model using the largest value.

## The final value used for the model was mtry = 35.

**varImp**(rf\_model)

## rf variable importance

##

## only 20 most important variables shown (out of 68)

##

## Overall

## DIABETE3Yes 100.00000

## DIABETE3No 38.60744

## PREDIAB1Not asked or Missing 2.17984

## DIABETE3No, pre‐diabetes or borderline diabetes 1.59779

## SSBSUGARNot asked or Missing 1.07324

## DIABETE3Refused 1.04852

## PDIABTSTNot asked or Missing 0.91942

## DIABETE3Yes, but female told only during pregnancy 0.80926

## SSBFRUT2Not asked or Missing 0.75418

## PREDIAB1No 0.56942

## SSBSUGARTimes per day 0.25372

## LIFECHGYes 0.22104

## LIFECHGNot asked or Missing 0.20559

## PREDIAB1Yes 0.18966

## PDIABTSTYes 0.16574

## PDIABTSTNo 0.07272

## EMPLOY1A student 0.04448

## SSBFRUT2Times per week 0.03808

## PREGEVERYes 0.03717

## PRNTLVIT0 times a week 0.03668

pred <‐ **predict**(rf\_model, newdata=test, type="raw")

**table**(pred,test$Diabetic\_Ind)

##

## pred 0 1

## 0 1791 1

## 1 0 268

© Prudential Financial