Using Machine Learning to Detect and Reconstruct Fake Good Responses to Big 5 Personality Questions: a Salesperson Job Interview Application

Derek Sweet
derekallen.sweet@studenti.unipd.it

Marina Vicini
marina.vicini@studenti.unipd.it

Emanuele Zangrando
emanuele.zangrando@studenti.unipd.it

Abstract

To measure the potential performance of a candidate during a job interview for a salesperson position, a personality test may be applied. The test considered is the Big 5 test, composed of 10 questions. However, as applicants apply, they may lie on the questionnaire to make themselves look more like a high performing employee. The aim of the report is to detect whether a person is lying on the questionnaire and in which questions s/he is doing so. Then, the final goal is to reconstruct the honest response.

Logistic Regression was the best model for discriminating honest and dishonest responses. It was favored over other models for its ease of interpreting and an accuracy score of 83%. Further discovery showed that discrimination was possible with just a single question as input ES1G - "gets nervous" to a Logistic Regression without losing significant predictability.

Furthermore, similar models have been implemented to do discrimination question by question. When comparing the models to TF-IDF, Random Forest was selected as the method that performed better when averaged over all questions.

For the final task, several models were built that successfully reconstructed the dishonest responses better than the trivial strategy. The best model was a CNN categorical autoencoder that has a reconstruction accuracy of 44.3% on average over all questions.

1. Introduction

It is human nature to exaggerate or sometimes even lie in order to appear in a way that will produce a favorable outcome. One common case when this dishonest behavior occurs is when applying or interviewing for jobs.

One method for measuring potential performance of an interviewee is the asses their personality based on the Big 5 personality traits. As applicants apply, they may lie on questions about personality traits to make themselves look like a high performing employee. The aim of this report is to assess ways to identify and reconstruct dishonest job interview responses for a salesperson position using both honest and dishonest responses collected through a Big 5 personality trait questionnaire.

This report will address the challenges, related research, relevant theory, along with experiments and an interpretation of their results. The experiments conducted in the paper showcase how to improve the existing machine learning modeling approaches to correcting dishonest to honest responses.

1.1. Challenges

The main goal of this project is to predict liars and infer their real answers by doing an "ad personam" analysis.
This kind of analysis could be interesting both from a practical and from a psychological point of view. More than being able to predict, we would like to get some insights on how people tend to lie in this kind of scenario. In order to achieve both goals we need to find a good balance between complexity and interpretability.
From a psychological point of view (if the model is not too complex) we can also try to get some information on what is the "typical human strategy" in lying to achieve the specific requested goal.

2. Dataset

The dataset is a questionnaire on the 5 dimensions of human personality focusing on job interviews for a salesperson. Each of the 230 respondents were asked to answer 10 questions about their personality honestly, then again dishonestly with context of a job interview for a salesperson. This "within subject" design allows for direct comparison and analysis of each individuals dishonest strategy.

Each of the Big 5 personality traits had two questions per personality trait and were on a 1-5 Likert scale. One of the two questions per personality trait were asked in reverse where the scale was inverted to remove bias. For the purposes of this report, all scores were converted to the same scale where 5 is high for the personality trait and 1 is low for the personality trait.

The questions asked were "I see myself as someone who ...":

- **EX1G** - is outgoing, sociable (extroversion)
- **EX2G** - is reserved (extroversion reversed)
- **A1G** - tends to find fault with others (agreeableness, reversed)
- **A2G** - is generally trusting (agreeableness)

- **C1G** - tends to be lazy (consciousness, reversed)
- **C2G** - does a thorough job (consciousness)
- **ES1G** - gets nervous easily (emotional stability, reversed)
- **ES2G** - is relaxed, handles stress well (emotional stability)
- **O1G** - has an active imagination (openness)
- **O2G** - has few artistic interests (openness, reversed)

3. Descriptive Analysis of Data

As a first step before building models, an initial descriptive analysis was conducted by aggregating responses to see general trends and make hypotheses.

Given the objective of the fakers, the expectation was that the fake good behaviour would generate higher scores than the honest responses. As seen in Figure [1], dishonest responses on average have higher score across all questions. The only questions that are not not *significantly* higher are EX2G - "is reserved" and C2G - "thorough job".

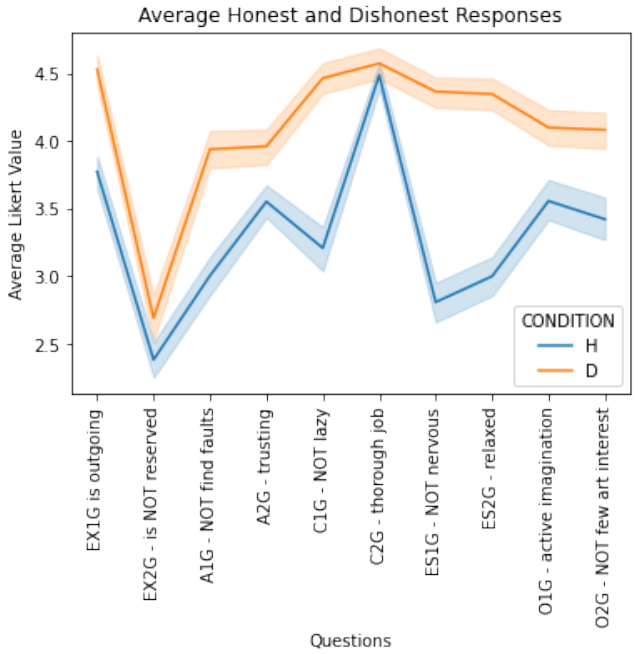

Figure 1. Average response by question separated by honest and dishonest responses

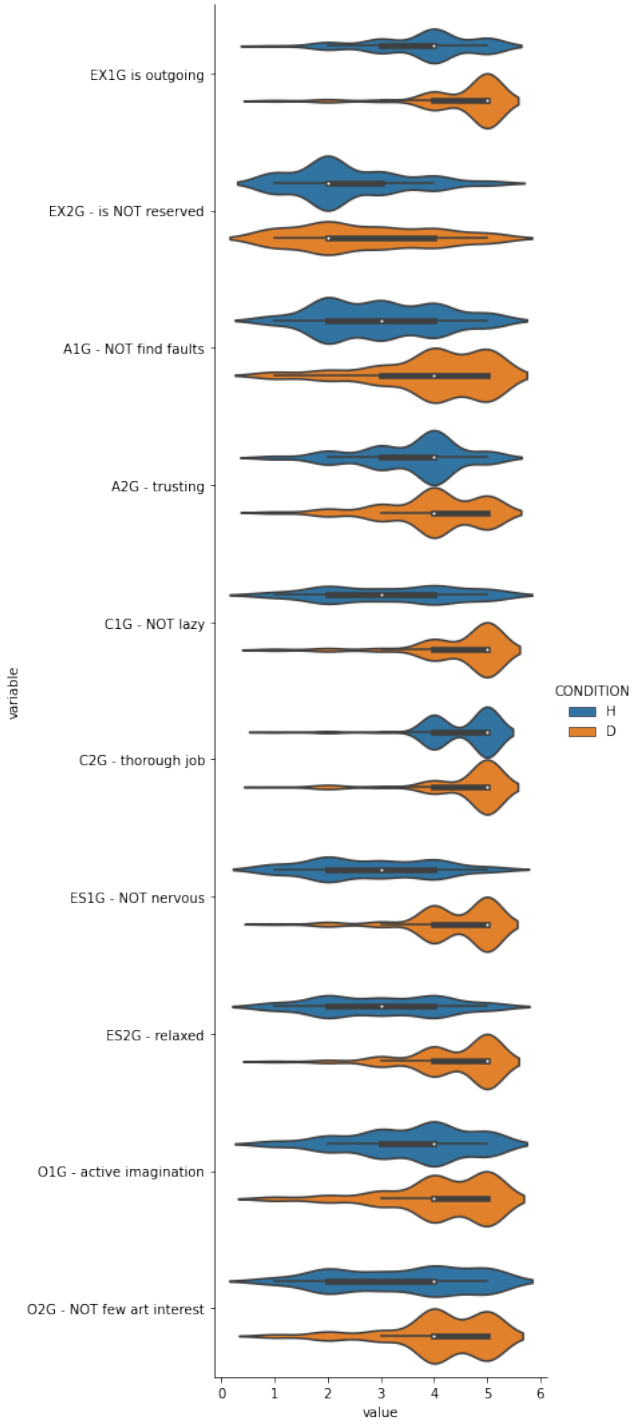

Figure 2. Distribution of response by question separated by honest and dishonest responses

3.1. Extroversion

EX2G - "is reserved" has a low score for both honest and dishonest responses. This result is counter intuitive given fakers will want to appear more extroverted. The fake good extroversion can be seen in the other extroversion question EX1G seen in Figure [1]. Honest scores have a higher concentration of responses around 4 while dishonest scores are more evenly distributed as seen in Figure [2].

3.2. Agreeableness

The response from A1G - "find faults" indicates fakers that want to appear they do not judge others. Honest responses are evenly distributed while fakers tend to concentrate their responses to 4 and 5, (Figure [2]). On the other hand, fakers tend to not lie as much with question A2G - "trusting" as they have very similar distributions.

3.3. Consciousness

Consciousness is often most associated to high performance on the job. Many dishonest responses fake that they had high Consciousness specifically to C1G - "lazy". C1G had the third largest difference between the honest and dishonest response.

As for the other consciousness question, C2G - "thorough job" appears as if everyone believes they are thorough re-

gardless if they are faking or not and both honest/dishonest give high scores.

3.4. Emotional Stability

Both questions for Emotional Stability have the largest difference between honest and dishonest responses. The honest distributions skew slightly lower indicating they are nervous and less relaxed. Most fakers go to the extreme to appear to handle pressure and stress well.

3.5. Openness

Openness questions have a slightly higher dishonest responses. Honest responses of O2G - "few artistic interest" have an even distribution, while fakers may be trying to seem more creative with higher responses of openness (Figure [2]).

3.6. Analysis of Faking Strategy

Participants are instructed to lie, but their lying strategy may differ. Figure [3] illustrates the number of faked questions and on average an applicant fakes 6.5 questions. Figure [4] shows the behavior of responding per question, whether applicants tend to increase, decrease or not change their response while retaking the questionnaire. In general, as expected, changed response tends to increase rather than decrease.

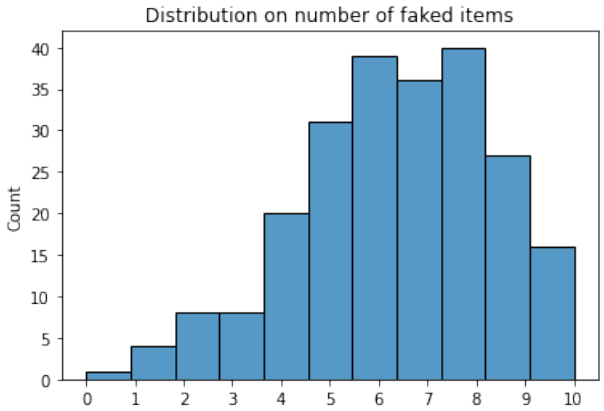

Figure 3. Amount of participants that faked that number of questions

When comparing the correlations between questions broken out by honest and dishonest responses, it is clear from Figure [5] that there is minimal correlation among honest question responses. The only questions that are somewhat correlated are ES1G and ES2G (0.61 correlation coefficient) along with O1G and O2G (0.42 correlation coefficient). These are both intuitive as they are both questions for Emotional Stability.

When people are dishonest, many of the questions become correlated indicating people may be oversimplifying the personality of a salesman (as seen in Figure [5]).

3.7. Literature about Big 5 Personality Traits

Literature provides evidence for the relationship between personality and job performance, substantiating the need to lie during a personality test in a job interview. In particular, research points at conscientiousness as the best indicator

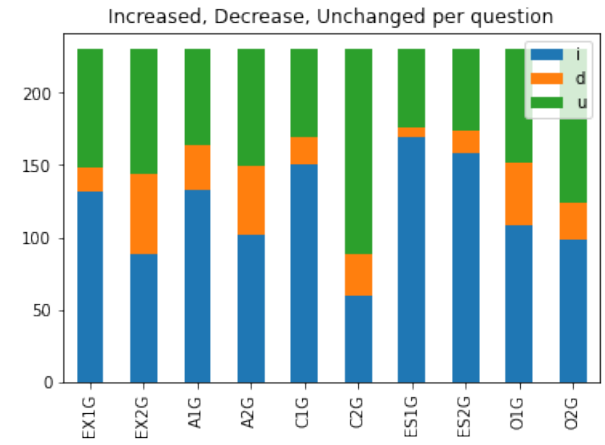

Figure 4. Amount of increased, decreased or unchanged responses per question.

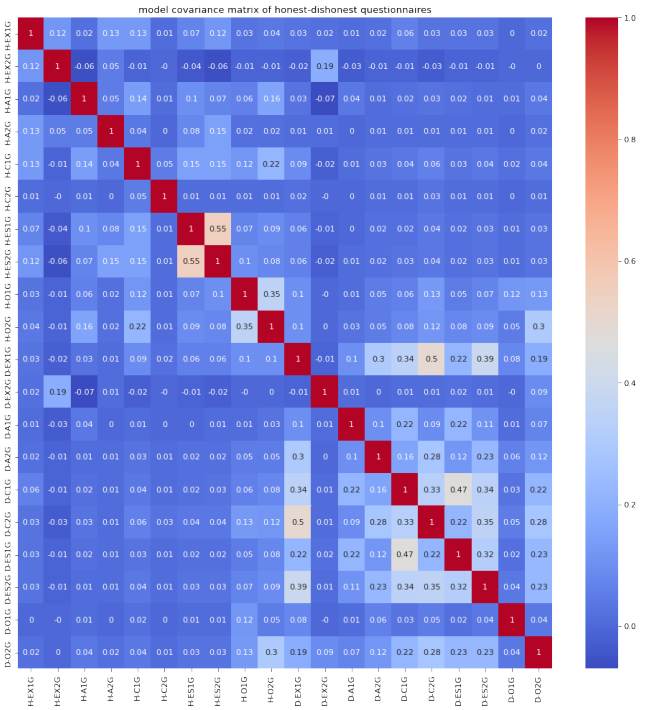

Figure 5. correlation matrix of joint honest dishonest questionnaires estimated by the Lasso graphical model.

for success [14]. Characteristics of people with high scores in conscientiousness are organized, plan oriented and determined. Regarding a sales context, people with high level of Extroversion, Conscientiousness and Emotional Stability. perform better [8], while no correlation was found with Agreeableness and Openness. These last two become important predictors respectively in team performance and creative tasks [2, 10].

4. Methods

In this section we will review several methods that are going to be used in the Experiment section. In the first four subsections, classification methods will be considered, while from subsection [4.5] the reconstruction methods are explained. In section [4.9], the measuring performance metrics are introduced.

4.1. Polynomial Feature Creation

Polynomial Feature Creation is a method used to expand the existing variables to include the interaction effects between features [**?**]. For the purposes of this paper, only second degree polynomials were used, meaning that every pairwise combination of questions were multiplied together. This method was used in hope to further differentiate the honest from the dishonest fake good responses.

4.2. Principle Component Analysis

Principle Component Analysis is dimensionality reduction technique that projects the original features into a new feature space. Each principle component maximizes the explained variance in the data [18]. For the purposes of this paper, PCA was used to see if new components could be created that have strong predictive power with less dimensions.

4.3. Discrimination Methods

There are many different models that can be used for binary classification of honest and dishonest responses. In this section there is a brief description of the methods used.

Logistic regression is a generalized linear model that leverages a fitted sigmoid curve for binary classification. A fitted logistic regression draws a linear decision boundary for classification and outputs a probability the observation belongs to the measured class [17]. One major advantage to using logistic regression is that it is interpretable. The coefficients learned by the model are the expected change in log odds given a 1 unit increase.

Linear Discriminant Analysis has a lot of similarities to PCA as both find linear combinations of features. While PCA is maximizing the explained variance through a linear combination of features, LDA aims to find the optimal linear decision boundary [16].

K-Nearest Neighbors classifies honest vs. dishonest responses based on the nearest neighbors as opposed to optimizing parameters using loss functions and gradient descent [1]. KNN accuracy highly depends on the quality and quantity of the training data with the ability to offset errors by adopting a larger k value comparing more neighbors [11].

Support Vector Classification is a supervised learning method, that works well because it optimizes the margin hyperplane that is defined by a subset of vectors that lie closest to it (support vectors). SVC belongs to linear classifiers, but it is strong at identifying non-linear classification problems using the kernel method such as the Radial Basis Function [6]. This improves the running time and minimizes error due to poor decision boundaries [9].

Random Forest is an ensemble of multiple decision trees. Each tree is trained on a (1) bootstrap bagged sample of observations and a (2) subset random sample of features which causes variance among the trees to ensure differing predictions ans less bias [19]. Where there is a majority vote among trees, the classification is made. One negative effect of the bootstrapping and random feature selection is that the model will become less interpretable.

XGBoost is a form of a parallelized gradient boosting decision tree that leverages an ensemble of weak predictors [4]. Any error produced in earlier trees will be boosted and predicted with more importance by the next weak learner [15]. Learning rate is a hyper parameter that shrinks the boosting process as a form of regularization. Given that decision trees are highly interpretable, this is a method that allows for interpretation of the importance of each feature in classification.

Multi-layer Perceptron finds non-linear relationships through non linear hidden activation units that propagate through a network of nodes until reaching a binary classifier sigmoid node. The flexible fitting mechanism guided by back propagation optimizes the MLP to work well with learning the non-linear relationships that might exist in classifying honest or dishonest responses.[5]

4.4. TF-IDF

On the data of our problem, a related work classified question by question whether it was faked or not through an information retrieval algorithm called Term Frequency-Inverse Document Frequency (TF-IDF). This method takes into account the answering style of each applicant [12]. For each participant and for each question, it computes a score based on how frequent the participant gave that response, and how frequent that response was given in that question. From these scores, an appropriate threshold is chosen to discriminate between honest and dishonest.

4.5. Denoising Autoencoders

Autoencoders are a special form of a neural network that contain an encoder to learn a latent representation of the input and a decoder to reconstruct the input from the latent space. The latent space is often a vector that is significantly smaller than the input and output, forcing the non linear model to learn the most representative features of the input and compress them into the latent vector. Then, this vector holds enough information for a decoder to learn how to reconstruct the original input. Training loss selection is computed depending on the desired output, but in the case of reconstructing Likert responses, mean squared error should be used after passing the output through a linear function.

Autoencoders have many applications, but for the task at hand, the denoising autoencoders have a particular use case. The denoising autoencoders will take as input a dishonest response to learn the latent space, but calculate the loss of the reconstructed response on the same person's honest response. This allows the latent space to learn a reconstruction manifold to change the dishonest response to the predicted honest response [**?**].

4.6. Graphical Model for the joint H-D distribution

Modeling the joint distribution of honest-dishonest was the first idea we exploited in order to infer the "lying strategy" from the data. We decided to use a Lasso graphical model to select only significant interactions in the joint distribution [7], as it can be seen in Figure [5]. This model helped us in realizing that probably the benchmark strategy is too simple, since interaction effects between the corresponding questions in the honest and dishonest version have a significant correlation only half of the times: this means that for half of the questions the direct effect on the same question answered honestly is not significant. A big advantage in using a graphical model on the joint distribution is that any kind

of inference can be done within the model, making it one of the most interpretable options.

4.7. Feature spreading

The next strategy we tried for reconstruction was feature spreading [3]. In particular for each person we constructed a vector of features containing his/her honest/dishonest questionnaires and an indicator vector telling in which questions he/she lied. When we have a new questionnaire, we start from the neutral hypothesis that each question has a 0.5 probability of being lied and that the honest and dishonest responses are not changing for this subject. After that we construct a KNN graph using all the nodes and we let features spread in the graph in a Pagerank manner. The results allow both to infer which question the person lied on and its honest responses. The disadvantage of this approach is that leads to not interpretable results.

4.8. Probabilistic CNN Denoising autoencoder

The "raw" denoising autoencoder presents some issues, in particular:

1. Using the mean square error as loss function is implicitly assuming that the output $x_H|x$ is normally distributed with mean $f_\theta(x)$ and covariance matrix $\sigma^2 I$. In particular this means that given a questionnaire, there is only one obvious way in which the person could have lied. We consider this assumption to not be realistic, since the questionnaire is pretty far from being a full gauge for the person's strategy;

2. This model does not allow for a careful uncertainty quantification on the reconstruction since the covariance matrix is assumed to be not dependent on x.

Even if the second problem can be solved by letting the covariance of $x_H|x$ to depend on x, the first one cannot be solved without changing approach. The solution we proposed was to allow a more flexible behaviour on the output distribution, not imposing a unimodal one. In particular we propose an architecture in which the output is made of a row stochastic matrix, whose entries (i, j) estimates:

$$\Big(P(\text{honest answer to question}\, i = j \,|x\,;\,\theta)\Big)_{i=1,..,10, j=1,,..,5}$$

To implement this strategy we did not use as input the raw questionnaires, but we used for training the per sample estimator of the second moment matrix xx^T to include second order interaction terms between the variables. This choice is motivated by the fact that probably the information contained in moments higher than one is more able to discriminate between honest dishonest and it can make a more careful reconstruction without having the network to learn this features by itself. The choice of a CNN architecture allows to keep the number of parameters low and to exploit local information in the second moment matrix.

To train it, we augmented the training set with Gaussian noise and we regularize it with elastic net on the first layers (to enforce even sparser interaction between the features) and dropout on the last layer.

4.9. Measuring Performance

4.9.1 Classification

Since the dataset was balanced, for each one of the classification models we used accuracy for evaluation. Moreover, we took under consideration the confusion matrix to analyze the results about false positive and negative rates.
All this evaluations have been done by using cross validation, to estimate also the variance of the metrics over folds.

4.9.2 Reconstruction

To measure reconstruction accuracy we decided to use estimators for the marginal correct reconstruction probabilities across questions (we constructed them by using cross validation for each model).

Again, the advantage of using this measure for reconstruction is that it naturally provides a confidence interval. Hence, a confusion matrix can be built for each one of the questions in order to study where each model fails.

5. Experiments

5.1. Amount of faking

The first experiment focuses on predicting the number of questions faked. First, a K-Nearest Neighbors was implemented obtaining an accuracy of 54.34% for predicting the number of unchanged questions and the same accuracy for predicting the number of increased questions. Then the same model was implemented taking as input the TF-IDF scores rather than the raw data. However, this did not improve the accuracy of predicting the number of unchanged question.

Since there is a low accuracy in predicting the number of faked responses, this variable was not considered as an input of future models.

5.2. Overall Discrimination

The first goal of the project is to be able to identify which questionnaire responses are honest and which are dishonest. To determine the best discrimination method to use, 10 different models were tested with 10 fold cross validation. Of the tested models:

- 7 of them (LogReg, LDA, KNN, SVC, RF, XGB and MLP) took as input the ten questions
- 1 of them (PolyLDA) took as input the polynomial interaction effects between the 10 questions (55 features in total)
- 2 of them (LR ES1G and LR PCA1) took as input a single feature for prediction

As seen in Figure [6], all the methods that take as input the 10 questions have relatively similar accuracy (82-85%). Logistic Regression, which is designed to find a linear decision boundary, gets 83% accuracy. Methods that are able to learn non-linear relationships (KNN, SVC, RF, XGB, and MLP) do not get significantly higher performance which is an indication that the discrimination between honest and dishonest responses is mostly linear in nature.

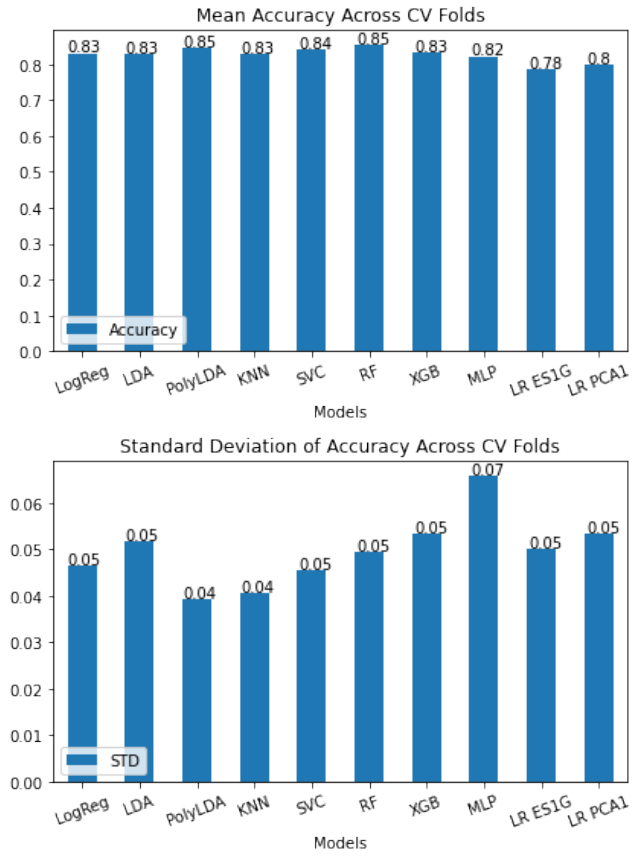

Figure 6. Plot of average and standard deviation of the overall discrimination accuracy across 10 cross validation folds

When interpreting the Logistic Regression coefficients (Figure [7]), there was one feature (ES1G) that was identified as the most important for prediction. A Logistic Regression was tested with only ES1G as input for discrimination to see the importance of this question in detecting fakers. As a result, ES1G can predict honest responses with 78% accuracy (as seen in Figure [6]).

Going one step further, PCA was performed to find more powerful features for prediction. A single principle component had the majority of explained variance in the data (35%). Using this single principle component, the model gets 80% accuracy (shown in Figure [6].

The most complex method tested was MLP which used a single hidden layer of 32 ReLU units with a 0.1 dropout. The variability of each fold was very large, likely due to the small sample size of the data. These models tend to perform better with large amounts of data.

A few patterns were identified in when inspecting the False Negative and False Positive responses from the Logistic Regression discrimination (Figure [8]). The responses that were classified as honest but were in fact dishonest closely align to the distribution of honest responses. A key driver of the false classification was that these fake responses for ES1G were higher than the honest responses, but significantly lower than the dishonest responses.

Conversely, the responses that were classified as dishonest but were in fact honest had very high Likert responses for nearly all questions. The honest responses look like they were dishonest because the most common strategy for dishonesty is to respond high for most questions.

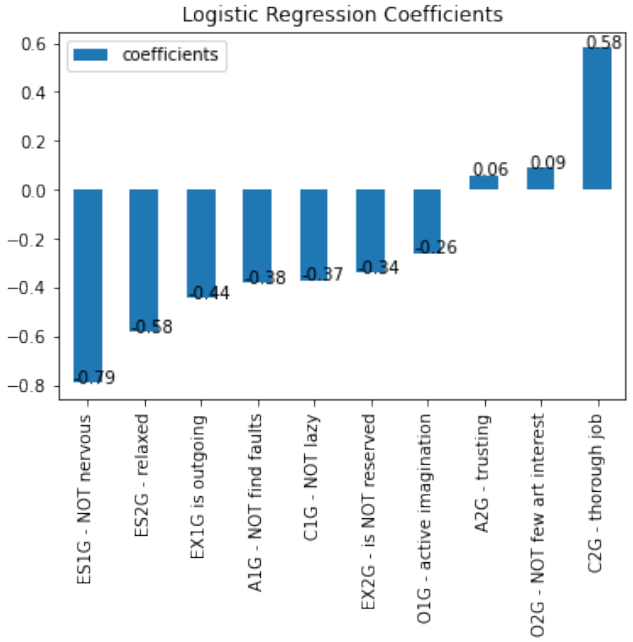

Figure 7. Plot of the feature coefficients in the logistic regression

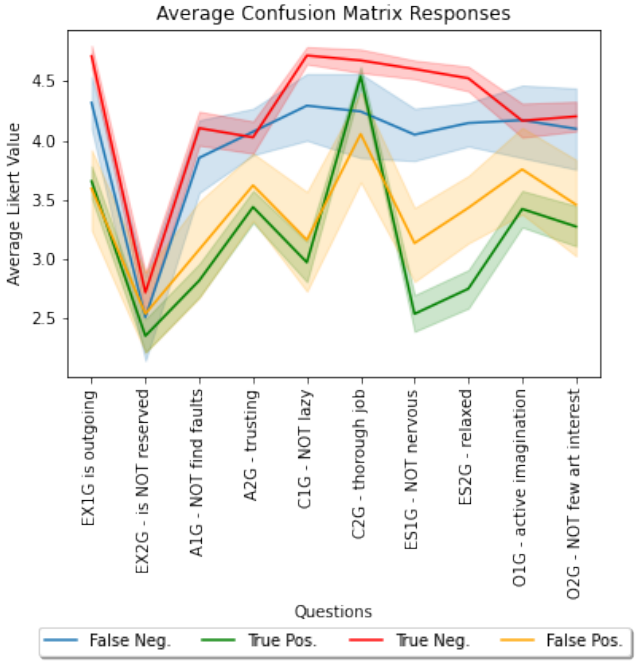

Figure 8. Plot of average responses by each of the 4 confusion matrix classifications

Another major reason for misclassification was due to respondents who had unusual dishonest strategies. The misclassified respondents on average changed 3.3 questions in the expected direction (fake good) which is -37% lower than all respondents. The strategy for the misclassified was to keep half of the responses unchanged: 4.9 unchanged responses on average which is 39% higher than all respondents.

Given the patterns found in the discrimination errors, one additional model was conducted to add interaction effects between the questions. This feature creation was implemented to try to capture more information about the individual strategy of each respondent. Given the high dimen-

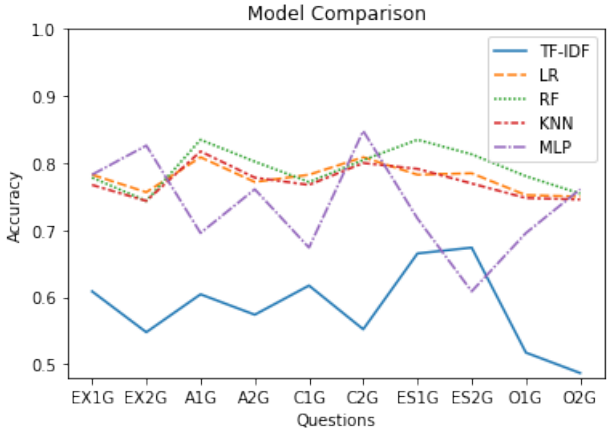

Figure 9. Comparison of accuracy per each question of models: TF-IDF, Logistic Regression, Random Forest, K-Nearest Neighbors and Multi Layer Perceptron.

sionality of the new feature space, LDA was used. The results improved from 83% in the original LDA to to 85% with the PolyLDA and the variance reduces from 5% to 4% (Figure [6]). PolyLDA had the best accuracy tied with Random Forest and lowest variance. Given the interaction effects are difficult to interpret and only provide a small improvement over the simpler models, it is recommended to favor more interpretable models such as the Logistic regression.

5.3. Discrimination by Question

Another goal of the project was classifying in each questionnaire response which answers where faked and which where not.

The first method used was the TF-IDF one in order to obtain a baseline. The average over the accuracy of each question is 58.48%. The model with the highest accuracy for the overall discrimination was Random Forest, so it was implemented for this task through a 10 fold cross validation. Other two models evaluated for this task were K-Nearest Neighbors (performing a 10 fold cross validation) and a Multi-Layer Perceptron. The results can be see in Figure [9]. The model with the worst performance is TF-IDF, while Random Forest was the best with a 79.28% average accuracy among the questions.

Others approaches that were attempted and did not improve performance are: (1) considering only the responses that have been classified as fake and (2) considering models that take into account balancing honest and dishonest responses through bootstrapping.

5.4. Reconstruction of Honest Response

Multiple models were compared for their ability to reconstruct responses as seen in Figures [13],[10].
In particular, we were interested in comparing the basic and the probabilistic denoising autoencoders. To do this, we cross validated the models and computed estimators for the top 1 along with top 2 accuracy for the probabilistic model. By using a probabilistic reconstructor we showed that the full distribution over labels gives us a lot more than a hard classification algorithm. We showed in fact that even though the top 1 accuracy of the probabilistic one is already better

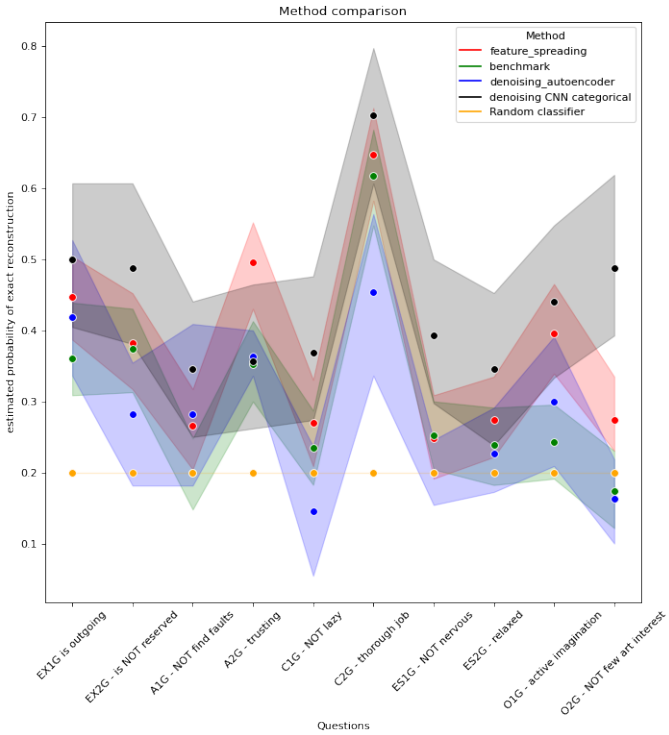

Figure 10. Estimated reconstruction accuracies per question for best performing models with their relative 95% bootstrapped confidence intevals.

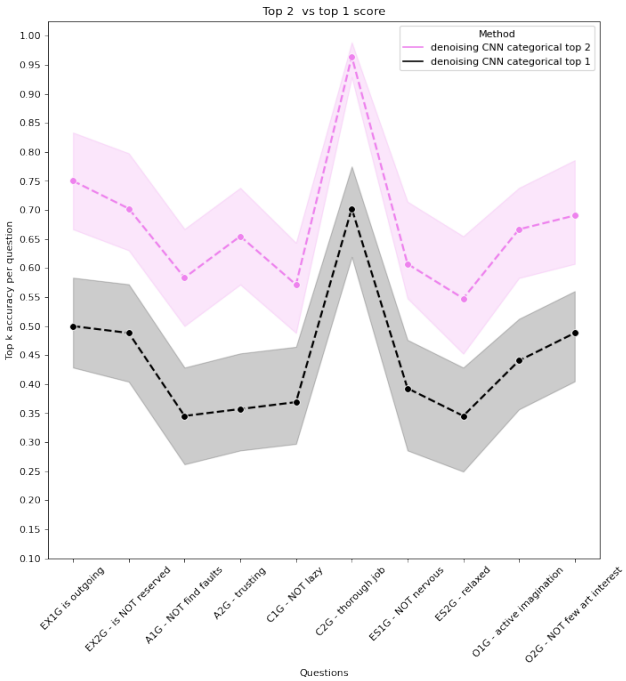

Figure 11. Top 1 and top 2 accuracies scores for the probabilistic autoencoder with their relative 95% bootstrapped confidence intevals.

than the raw one, top 2 accuracy is significantly increasing the performance as shown in Figure [11].
This simple experiment was meant to show the advantage of having a full probability distribution as an output: by looking at it we can assest how much the reconstruction we are proposing is likely to be the right one according to the model. Average accuracy over questions are presented in Figure

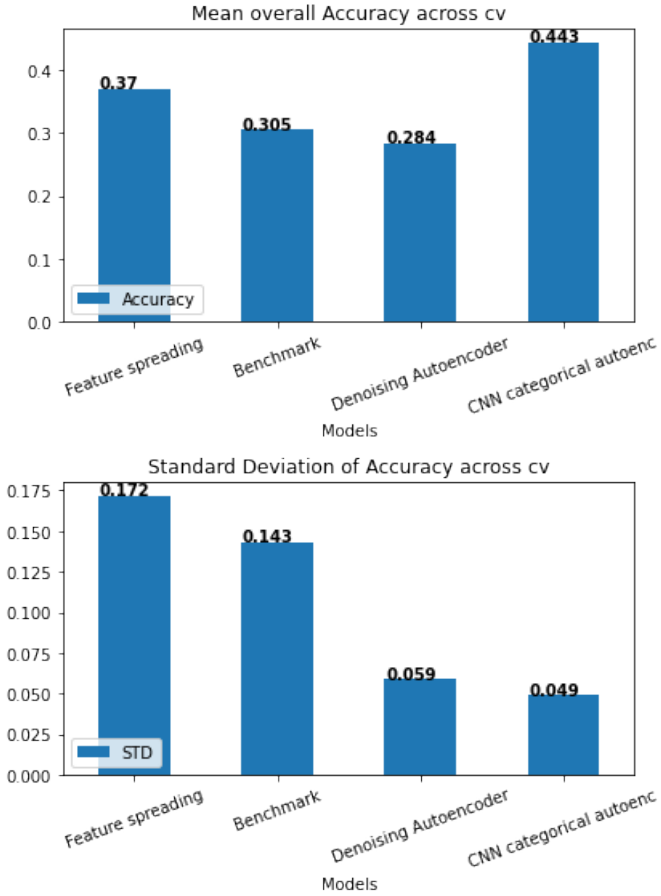

Figure 12. Average estimated probabilities of correct reconstruction over all questions with their standard deviations.

[12], where we can see that the probabilistic CNN architecture is leading in both performance (44.3%) and variance of the results, immediately followed by feature spreading (37%).
An interesting further investigation could be to explore Bayesian versions of this model to assess epistemic uncertainty in the parameters [13].

6. Conclusion

It can be seen though the experiments that fake responses can be detected with strong accuracy and reconstructed better than a trivial benchmark.

During a job interview for the position of a salesperson, the candidate may tend to fake good in a personality test in order to look like a high performing employer. The literature about Big 5 personality tests and job performance indicates Consciousness as the best indicator for success. The expectation was that participants would fake good the most on Consciousness questions, however our experiments show that Emotional Stability (specifically nervousness) was a more important indicator to detect lies. Nervousness was a key predictor because participants appear to give very realistic and low answers when responding honestly, but have very high responses when faking. Conversely, the question about doing a "thorough job" was a weak predictor as the honest results were skewed higher due to social desirability.

The confidence of detecting a response as fake is 83%,

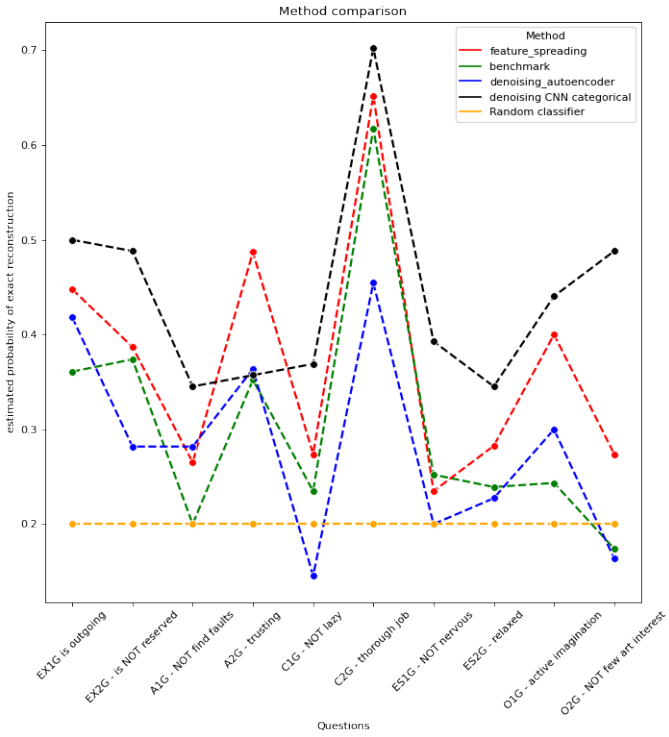

Figure 13. Per question estimated reconstruction probability for all the models under consideration.

while the average of accuracy to detect each question as honest or dishonest is 79%. Hence, we recommend reconstructing the whole honest response instead of just certain questions. The accuracy for the reconstruction of the honest response is 44.3% on average over all questions, obtained through a CNN categorical autoencoder. Moreover, this score is significantly better than the trivial reconstruction and also the basic denoising autoencoder for almost all questions.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Bret H Bradley, John E Baur, Christopher G Banford, and Bennett E Postlethwaite. Team players and collective performance: How agreeableness affects team performance over time. *Small Group Research*, 44(6):680–711, 2013.

[3] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[4] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.

[5] Giuseppe Ciaburro, Balaji Venkateswaran, and Li hong cheng. *Shen jing wang luo: R yu yan shi xian = Neural networks with R*. Ji xie gong ye chu ban she, 2018.

[6] Paolo Giudici. *Data Mining. Modelli informatici, statistici e applicazioni)*. McGraw-Hill, 2005.

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman Hall/CRC, 2015.

[8] Andreas Klang. The relationship between personality and job performance in sales:: A replication of past research and an extension to a swedish context, 2012.

[9] Yoonkyung Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99:67 – 81, 2004.

[10] Hector Madrid and Malcolm Patterson. Creativity at work as a joint function between openness to experience, need for cognition and organizational fairness. *Learning and Individual Differences*, 51, 08 2015.

[11] Stefanos Ougiaroglou and Georgios Evangelidis. Dealing with noisy data in the context of k-nn classification. In *Proceedings of the 7th Balkan Conference on Informatics Conference*, BCI '15, New York, NY, USA, 2015. Association for Computing Machinery.

[12] Alberto Purpora, Dora Giorgianni, Graziella Orrù, Giulia Melis, and Giuseppe Sartori. Identifying faking to single item responses in personality tests: a new tf-idf based method. 2021.

[13] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, Inc., USA, 1st edition, 2015.

[14] Michelangelo Vianello, Egidio Robusto, and Pasquale Anselmi. Implicit conscientiousness predicts academic performance. *Personality and Individual Differences*, 48(4):452–457, 2010.

[15] Wikipedia contributors. Gradient boosting — Wikipedia, the free encyclopedia, 2021. [Online; accessed 4-December-2021].

[16] Wikipedia contributors. Linear discriminant analysis — Wikipedia, the free encyclopedia, 2021. [Online; accessed 4-December-2021].

[17] Wikipedia contributors. Logistic regression — Wikipedia, the free encyclopedia, 2021. [Online; accessed 3-December-2021].

[18] Wikipedia contributors. Principal component analysis — Wikipedia, the free encyclopedia, 2021. [Online; accessed 4-December-2021].

[19] Wikipedia contributors. Random forest — Wikipedia, the free encyclopedia, 2021. [Online; accessed 4-December-2021].