The Law of Equal Opportunities or Unintended Consequences? The impact of unisex risk assessment in consumer credit

Galina Andreeva

University of Edinburgh, UK

and Anna Matuszyk

New York University, USA

Warsaw School of Economics, Poland

Summary. Gender is prohibited from use in decision-making in many countries. This does not necessarily benefit females, as this paper shows by analysing a unique proprietary dataset on car loans from an EU bank. The results suggest that inclusion of Gender as a dummy variable is statistically significant, but does not alter the predictive accuracy of the model. Yet the proportions of accepted women/men depend on whether Gender is included. The paper explores the association between predictors in the model with Gender, to demonstrate the omitted variable bias and how other variables proxy for Gender. It points to inconsistencies of the existing regulations in the context of automated decision-making.

Keywords: credit scoring; gender; decision-making

Version: February 7, 2019

1 Address for correspondence: Dr Galina Andreeva, Credit Research Centre, University of Edinburgh Business School, UK, E-mail: Galina.Andreeva@ed.ac.uk
1. Introduction

The paper investigates the consequences of restrictions on information in automated decision-making. It analyses an example from the area of quantitative risk assessment in retail credit, aka credit scoring, that is used to decide which applicants should be granted credit. Nevertheless, the results can be extended to situations of algorithmic decisions, in general, when the predictive algorithms are developed on historic data. Credit scoring is a collection of mathematical and statistical models that predict the probability of a borrower’s default, using historic data that may include personal characteristics such as age, income or residential status. A number of experts in credit scoring have suggested that large banks, retailers and insurers almost exclusively use automated credit scoring systems to decide whether to accept or reject an application (Anderson, 2017, Thomas et al., 2017). Already in 1997 Hand and Henley wrote: “Nowadays it seems that the only organizations which do not use credit scoring approaches are the smaller and/or more personal companies, and those concerned with corporate finance, where statistical methods have been slower to be adopted.” (Hand and Henley, 1997, p.531). There is also evidence that even smaller credit institutions, such as microlenders, are adopting this lending technology (Schreiner, 2002).

The paper uses a specific example when *Gender* cannot be included into a credit scoring model as a factor/predictor, and illustrates the consequences for a lender, and for male and female credit applicants. *Gender* is prohibited by law from use in decision-making in the majority of developed countries. The prohibition follows from anti-discrimination provisions, e.g. the European Equal Treatment in Goods and Services Directive (EU Council, 2004). There has been a long debate as to how the “equal treatment” should be applied in retail financial services. Initially the Directive included a special clause, allowing the use of *Gender* as a factor in risk assessment in insurance and related financial services, provided that it was justified by "relevant and accurate actuarial and statistical data" (Article 5(2), EU Council, 2004). Yet in March 2011, the European Court of Justice (ECJ) cancelled this provision (European Parliament, 2011). The insurance industry expressed the concern that since *Gender* is associated with risk, its removal will lead to higher motor premiums for women, who are known to be safer drivers compared to men (Oxera, 2010). The expectation was that the ECJ ruling would make premiums the same irrespective of *Gender*, yet McDonald (2015) showed that the difference in premiums between male and female drivers still remained as a result of ‘proxies’ – variables which are legal to use and are correlated with both *Gender* and risk, e.g. the customer’s profession. This correlation and the resulting insensitivity of the model’s
predictive power to simple removal of the prohibited variable has been noted before (Andreeva et al., 2004, Hand, 2012a, b). However, there is no clear guidance, as to what would be a discrimination-free solution.

In this paper we would like to illustrate empirically the consequences of legal restrictions on information in application to automatic decision-making, in order to highlight potential inconsistencies of the existing regulations and to inspire further research into better solutions. In contrast to McDonald (2015), we do not aim at evaluating the effect of a particular piece of legislation, but rather wish to illustrate how the principle of ‘equal treatment’ does not translate into ‘equal outcome’.

We do it by analysing a unique proprietary dataset on car loans from an EU bank, which contains Gender, other application characteristics and observed credit performance. Our illustration consists in following a standard credit scoring methodology that is used by banks in practice to construct a model based on credit application variables, including Gender. We then remove Gender (which is statistically significant) and comment on the changes in parameter estimates. We find that the predictive accuracy of the model is not affected, but when we simulate different scenarios for accept/reject decisions using models with and without Gender, we observe differences in proportions of men and women rejected by different models. Furthermore, we apply Bayesian Networks (BN) and Multiple Correspondence Analysis (MCA) to investigate the structure of associations between Gender and other predictors in the model. These associations can proxy for Gender, and in order to consistently follow the principle of equal treatment the correlated variables should be excluded from the model in addition to Gender. However, when these additional variables are removed, the significant deterioration in predictive accuracy is observed.

The paper contributes to the research and discussion on discrimination in several ways. First, we provide the empirical confirmation of Gender’s statistical significance in predicting the probability of default. Second, we demonstrate empirically the ineffectiveness of existing anti-discriminatory regulations that prohibit the use of Gender as a predictor in automated decision-making. As noted above, removing Gender from the model leaves the predictive accuracy practically unaffected, which would imply that for lenders there is little difference in terms which model should be used, since they can achieve similar level of classification performance. And yet for consumers it matters which model is used as a decision tool when accepting for credit. We show that female borrowers have relative higher acceptance rates when the model with Gender is used.
Third, we highlight the importance of the minority status, i.e. when the protected class constitutes a smaller segment in overall portfolio/population, it may not be represented appropriately if the distinction between the minority and majority class is not allowed. We use the current analysis as an example to demonstrate what happens to a minority class (women in our case) when this class possesses better qualities as compared to the majority. We believe that the intention of the law should be to protect disadvantaged and/or minority groups. In our case women are better credit risks as compared to the majority (men). By prohibiting the use of Gender, the law effectively removes the possibility for women to signal their quality and for the lenders to include this signal (Gender factor) into the screening process.

Our last (but not least) contribution consists in revealing the associations and dependency structure between all variables in the model, thus accentuating the inter-connected nature of this world and the difficulty of eliminating the effect of one single variable/factor.

The relevance of our investigation goes beyond the example of car loans, and we believe that our arguments apply to a wide range of situations where models are built on historic data, and the model captures the relations present there, which may include associations of protected groups with perfectly legal variables, such as profession. The model is then applied to predict the future performance, thus extrapolating historic relations into the future, and the decisions are made based on those predictions. The timeliness and importance of such investigation is demonstrated by the recent ‘Algorithms in decision-making’ inquiry of House of Commons Select Committee, where one of the questions was: “the scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome…” (Science and Technology Committee, 2017). The submission from The Royal Statistical Society noted the experience from credit scoring as the example worth of consideration (RSS, 2017).

The remaining of the paper is structured as follows. The next section briefly reviews the relevant economic theories of discrimination and previous empirical research. Section 3 gives an overview of the credit scoring methodology. Subsequent Section 4 describes the data and the results of the empirical analysis. The final section discusses the implications and concludes.
2. Anti-discrimination law, theories of discrimination and empirical research

In general, the law distinguishes between direct and indirect discrimination, although the details may differ across the countries and legislations. Direct discrimination is based on the principle of equal treatment, and is unlawful. It arises when a person is treated ‘less favourably’ because of the prohibited characteristic, such as gender or race. Indirect discrimination arises when a neutral rule or criterion is applied to everyone, but people with protected characteristics are disadvantaged. In the latter case the focus is on equal outcome, and such discrimination can be justified, e.g. by business necessity (EU Council, 2004).

Economic theory distinguishes between subjective taste-based discrimination (Becker, 1971) which is a consequence of subjective preferences or prejudices, and objective statistical discrimination (Phelps, 1972; Stiglitz, 1993). Statistical discrimination is the consequence of insufficient information necessary to estimate the level of risk. Some relevant information (e.g., intention to repay) cannot be observed, so if group membership is correlated/associated with latent information and both are associated with default, group membership can be used as a signal or a proxy. This distinction (subjective v statistical) is not reflected in the law. As noted by Alan Turing Institute (ATI, 2017), the anti-discrimination legal principles were developed in the era of human judgment, and do not cover certain aspects of automated decision-making. One of such aspects or inconsistencies has been pointed out by Hand (2009, 2012a) and is linked to the concept of statistical or group membership discrimination – the law denounces stereotyping based on group membership and calls for the individual approach, yet probabilistic models that underlie the majority of automated decision-making rely on summary estimates derived from aggregated data, i.e. a probability cannot be estimated based on a single individual observation.

Previous research has focused mainly on ways of identifying whether discrimination is occurring. One of the most used approaches is a regression-based methodology where the coefficient for group membership is taken as a measure of discrimination. There is a vast body of literature dedicated to this topic, which is not reviewed here, but the interested reader can refer to Andreeva *et al.* (2004) for the summary of this research and to Charles and Guryan (2011) for a detailed discussion of limitations of regression and other approaches to detect discrimination. The most wide-spread criticism refers to omitted variable bias, i.e. if group membership is correlated with residuals (unobserved or omitted variables), its coefficient will be biased, so it cannot be interpreted as an indication of discrimination.
Another popular strand of literature investigates if protected characteristics have negative or positive association with risk. Women have been found to be more reliable payers in a number of studies in a variety of contexts and credit products (e.g. Agarwal et al., 2016; D'Espallier et al., 2011, Do and Paley, 2013, Coin, 2013). We do not review this topic in detail either, because the objective of this paper is not to prove that women are better credit risks or that lenders engage in discrimination. Instead we would like to investigate empirically the effects of legal restrictions on information in automated decision-making and the consequences for a decision-maker and the segments that are represented by a prohibited variable (men and women).

Recent regulatory inquiries into algorithmic decision-making inspired research into discrimination arising from machine-learning. Fuster et al. (2017) show that more advanced algorithms, such as random forests, increase discrimination, as compared to a standard logistic regression. This is logical, since higher predictive accuracy means higher discrimination in a statistical sense (i.e. better separation between the risk classes) and if there are true differences in risk levels of protected groups, more accurate model will intensify this distinction. Berk et al. (2017) quantify this relationship as ‘Price of Fairness’, and show that when the predictive accuracy goes up, the ‘fairness’ will go down. Nevertheless, these papers do not analyse the effect of ‘equal treatment’ on ‘unequal outcome’, which is the main focus of this paper.

A number of studies (Fair, 1979, Johnson, 2004, Andreeva et al., 2004, Chan and Seow, 2013) have postulated certain implications from the prohibition, but have not provided any empirical proof. One implication is that if a prohibited variable is associated with default, its removal should lead to a reduction in predictive power, which should negatively impact on lenders through increased delinquency and cost of credit. It may also negatively impact on consumers, if increased cost of credit is passed on to them. This negative impact could be justified by benefits to protected classes, i.e. better access to credit, but whether in fact the prohibition increases chances of protected groups of being accepted for credit can also be questioned.

Another potential implication consists in restricted chances of being accepted for credit, and this can affect protected and unprotected groups. However, the investigation of these implications remains largely speculative in absence of suitable empirical data which are not available (except for limited information that is reported for mortgages in the US under Home Mortgage Disclosure Act (HMDA, 2015). Taylor (2011) notes lack of data as a major obstacle in discrimination research.
in non-mortgage credit. That is why, having obtained the relevant data, we would like to investigate the predictive value of Gender and its impact on access to credit for men and women.

3. Credit scoring methodology

The project follows the standard methodology for building credit scoring models as described in Anderson, 2017, Hand and Henley, 1997, Siddiqi, 2006, Thomas et al., 2017. This allows mimicking the process of credit model construction in practice, so that observations can be made about potential impact in credit-granting environment.

The authors above assert that logistic regression is the most popular and widely used algorithm in credit scoring and it is also used in this paper:

\[
\text{logit} (p_i) = \beta^T x_i
\]

where \( p_i \) is the probability of experiencing default (according to a selected definition) for customer \( i \) and \( x_i \) are predictor variables/ characteristics.

The model building process was preceded by the data transformation or coarse-classification, which is a standard approach in credit scoring (Jung and Thomas, 2008; Baesens et al., 2009). In case of continuous variables, the first step is to split a characteristic into intervals, usually between 10 and 20. In the next step, adjacent intervals with the similar default rates are merged into larger coarse classes. This allows to preserve any non-monotonic patterns, to treat outliers and to include missing values that become a separate coarse-class. Similarly, for categorical variables, small categories are grouped together in order to achieve more stable predictions. Finally, the resulting coarse-classes are either replaced by weights of evidence or transformed into binary dummy variables (Lin et al., 2012). In this paper we followed the dummy variable approach.

The estimated probability of default (PD) is used as a ‘score’, which can be viewed as a summary of credit worthiness. Credit applicants can be ranked on the basis of the score or in other words, according to the level of their attractiveness to lender. The accept/reject decision is achieved by setting a threshold or cut-off: customers with higher probability of default than the cut-off are rejected, whilst those with lower PD are accepted for credit.

In addition to standard measures of model fit, credit scoring models are evaluated in terms of their ability to discriminate between ‘good’ and ‘bad’ credit risk. Area under the Curve (AUC) is a common measure for the discriminatory power. The ROC (Receiver Operating Characteristics)
curve is obtained by plotting Sensitivity against 1-Specificity for various cut-off points of PD, with Sensitivity and 1-Specificity being defined in the following way:

\[
\text{Sensitivity} = \frac{TPOS(s)}{n_1},
\]

\[
1 - \text{Specificity} = 1 - \frac{TNEG(s)}{n_1} = \frac{FPOS(s)}{n_0},
\]

where:

- \(\frac{n_1}{n_0}\) are the numbers of events/defaults (marked as 1) and non-events (marked as 0) respectively;
- \(TPOS\) is the number of correctly predicted events; \(FPOS\) is the number of incorrectly predicted non-events;
- \(TNEG\) is the number of correctly predicted non-events;
- \(s\) is the cut-off used to classify predicted probabilities into events/non-events (\(p_i \geq s \rightarrow \text{event}; p_i < s \rightarrow \text{non-event}\)).

In the credit scoring context, Sensitivity can be interpreted as a cumulative proportion of defaults or the ‘Bad’ customers with and above a score \(s\) (therefore, correctly rejected) and 1-Specificity - as a cumulative proportion of non-defaults or the ‘Good’ customers incorrectly rejected (Thomas et al., 2017). The higher values of AUC would indicate superior models, no discrimination would correspond to AUC=0.5. This measure summarises the ability of the model to rank risk correctly over the whole range of possible scores or cut-offs, therefore does not require the choice of a specific cut-off.

It was shown by Hanley and McNeil (1982) that conceptually AUC corresponds to the Wilcoxon or Mann-Whitney statistic, which estimates the probability that a predicted PD of a randomly selected Bad account will be higher than or equal to that of a randomly selected Good account.

### 4. Data description and empirical results

The dataset is a portfolio of car loans issued from 2003 to 2009 by a major bank (which chose to remain anonymous) operating in an EU country. Table 1 summarizes the training sample, which is used for the model estimation; and test sample, which is reserved for assessing the model’s predictive accuracy. Splitting the data into training and test samples is a standard methodology in credit scoring, here the split is 80% : 20%. ‘Bad’ are customers who missed two consecutive monthly payments – the definition used by the lender that provided the data.


Table 1. Frequencies and percentages for men and women in training and test samples.

|                      | Training sample | Test sample |
|----------------------|-----------------|-------------|
|                      | Good | Bad | Total | Good | Bad | Total |
| # of female customers| 16746| 220 | 16966 | 4186 | 55  | 4241  |
| % of female customers| 98.70% | 1.30% | 26.71% | 98.70% | 1.30% | 26.71% |
| # of male customers  | 45696| 847 | 46543 | 11424| 212 | 11636 |
| % of male customers  | 98.18% | 1.82% | 73.29% | 98.18% | 1.82% | 73.29% |
| Total # of customers | 62442| 1067| 63509 | 15610| 267 | 15877 |
| Total % of customers | 98.32% | 1.68% | 100%   | 98.32% | 1.68% | 100%   |

As can be seen from Table 1, females have lower proportion of Bad accounts as compared to males. It should also be noted that women are in minority, constituting slightly more than a quarter of the sample. For this reason and because the event to be modelled constitutes only 1.3% for women, and 1.82% for men, we used a random sampling stratified on Gender and Good/Bad status when dividing the data into training and testing sets. This is an accepted practice in predictive modelling, in particular, of rare events, when it is important that both samples are representative of the population(s) of interest. Kohavi (1995) has shown that a stratified random cross-validation is superior in terms of lower variance and bias as compared to a simple random one. A single split into a train/test set can be viewed as one iteration of a cross-validation with k=2 folds.

Four logistic regression models have been built:
1. Model with Gender (training sample comprising both men and women)
2. Model without Gender
3. Model for men only (training sample consisting of men only)
4. Model for women only (training sample consisting of women only).

We have decided to build segmented Models 3 and 4 as opposed to including interactions terms, because this approach is preferred in practice (Banasik et al., 1996, Bijak and Thomas, 2012, Thomas et al., 2017). This is also in line with our desire to demonstrate the possibility of achieving an ‘equal outcome’: with segmented models it is easy to accept/reject equal proportions of men and women. Later in Section 4.2 we explore the association between all variables used in modelling, including Gender.
4.1. Parameter estimates, predictive accuracy and reject rates

The results from the four models are reported in Table 2. We have used two criteria to include a variable into Model 1: the variable has to be statistically significant at 0.05; and it has to show a high predictive power in explaining the PD. We measured the predictive power by AUC for each variable separately and selected 12 variables with the highest values. Variables that have been selected into Model 1 with Gender are retained in other models to allow for comparisons of parameter estimates.

As explained in Section 3, we followed the established practice in credit scoring of coarse-classification (or categorisation), and used dummy variables to represent the final categories. The reference category is selected to be the largest one.

Table 2. Parameter estimates (standard errors are in brackets) and model fit statistics for four logistic regression models to predict the PD. The reference category is given in brackets under the corresponding variable name. The significance level of the covariates is represented by the following symbols: † p-value <0.0001, ‡ p-value <0.005, § p-value <0.05.

| Variable                  | Attribute/category | % in category | Model with Gender (Model 1) | Model without Gender (Model 2) | Model for men only (Model 3) | Model for women only (Model 4) |
|---------------------------|--------------------|---------------|-----------------------------|-------------------------------|----------------------------|-------------------------------|
| Intercept                 |                    |               | -7.3942 † (0.1722)          | -7.5207† (0.1708)             | -7.6844† (0.2073)           | -7.0066† (0.3135)             |
| Gender Female             |                   |               | -0.457 † (0.0867)           |                               |                            |                               |
| 1 kid                     |                   | 23.26         | 0.19 (0.1009)               | 0.1525 (0.1000)              | 0.267§ (0.1219)            | 0.1248 (0.1874)              |
| 2 kids                    |                   | 15.04         | 0.1918 (0.1302)             | 0.1763 (0.1298)              | 0.2192 (0.1552)           | 0.25 (0.2447)                |
| 3+ kids                   |                   | 3.12          | 0.3553 (0.2313)             | 0.3494 (0.2310)              | 0.1584 (0.2950)           | 1.1087‡ (0.3783)            |
| Missing information       |                   | 10.87         | -0.6816 † (0.1254)          | -0.6944† (0.1251)            | -0.7207† (0.1440)         | -0.5017§ (0.2618)           |
| Car price                 |                    |               |                             |                               |                            |                               |
| Cheap                     |                   | 5.28          | -1.0987† (0.1326)           | -1.1048† (0.1322)            | -1.2304† (0.1548)         | -0.6552§ (0.2718)           |
| Medium price higher       |                   | 59.58         | 0.426‡ (0.1099)             | 0.4406‡ (0.1095)             | 0.3752‡ (0.1271)          | 0.5243§ (0.2277)           |
| Expensive                 |                   | 15.87         | 1.1813† (0.1116)            | 1.1955† (0.1112)             | 1.2106† (0.1286)          | 1.0349‡ (0.2378)            |
| Down payment, %           |                    |               |                             |                               |                            |                               |
| (ref. (35%, 50%))         | <=25%              | 16.87         | 1.2702† (0.1087)            | 1.2603† (0.1085)             | 1.3522† (0.1260)          | 0.98† (0.2199)              |
|                          | (25%,35%)         | 8.65          | 0.7133‡ (0.1248)            | 0.7096‡ (0.1246)             | 0.7704‡ (0.1443)          | 0.515§ (0.2555)             |
|                          | 51%+               | 34.49         | -1.2147† (0.1940)           | -1.2075† (0.1941)            | -1.2145‡ (0.2268)         | -1.1765‡ (0.3766)           |

Electronic copy available at: https://ssrn.com/abstract=3212702
### Car age, years (ref: [0, 2])

| Car age, years | 2  | 3.25 | 4+  |
|----------------|----|------|-----|
| 1.56           | 1.311† | 1.842† | 2.530† |
|                | (0.1454) | (0.1196) | (0.1348) |
| 1.3197†        | 1.8691† | 2.5635† | 2.4991† |
|                | (0.1448) | (0.1191) | (0.1343) |
| 1.3112†        | 1.9397† | 2.4991† | 2.6423† |
|                | (0.1678) | (0.1362) | (0.1558) |
| 1.3781†        | 1.5436† | 1.6423† | 1.6055† |
|                | (0.3027) | (0.2748) | (0.2853) |

### Loan duration, months (ref: ≤36)

| Loan duration, months | ≤36 | 60+ |
|-----------------------|-----|-----|
| 17.93                 | 0.7961† | 1.6451† |
|                        | (0.1181) | (0.1164) |
| 0.8012†               | 1.6554† | 1.9397† |
|                        | (0.1179) | (0.1162) |
| 0.8606†               | 1.7494† | 2.3574† |
|                        | (0.1387) | (0.1241) |
| 0.586§                | 0.678‡  | 0.684‡  |
|                        | (0.2320) | (0.1949) |

### Time in employment, years (ref. ≥7)

| Time in employment, years | <1 | (1;4] | (4;7] |
|---------------------------|----|-------|-------|
| 11.16                     | -0.4276† | -0.4551† | -0.4318† |
|                           | (0.1004) | (0.0945) | (0.0940) |
| 19.77                     | -0.1263† | -0.161   | -0.4511† |
|                           | (0.1289) | (0.1080) | (0.1107) |
| 25.06                     | -0.1053  | -0.1743  | -0.5646† |
|                           | (0.1538) | (0.1077) | (0.1083) |
| 17.08                     | -0.2415  | -0.2504§ | -0.5465§ |
|                           | (0.1817) | (0.1270) | (0.1188) |
| 17.08                     | -0.4565§ | -0.2504§ | -0.5465§ |
|                           | (0.1817) | (0.1270) | (0.1188) |
| 16.38                     | -0.1263† | -0.1743  | -0.5465§ |
|                           | (0.1538) | (0.1077) | (0.1083) |
| 2.99                      | -0.1053  | -0.2415  | -0.5465§ |
|                           | (0.1538) | (0.1270) | (0.1188) |
| 16.90                     | -0.2415  | -0.2504§ | -0.5465§ |
|                           | (0.1538) | (0.1270) | (0.1188) |
| 14.90                     | -0.1053  | -0.2415  | -0.5465§ |
|                           | (0.1538) | (0.1270) | (0.1188) |
| 23.39                     | -0.1263† | -0.1743  | -0.5465§ |
|                           | (0.1538) | (0.1077) | (0.1083) |
| 2.32                      | -0.1263† | -0.1743  | -0.5465§ |
|                           | (0.1538) | (0.1077) | (0.1083) |

### Net income (ref. Medium income higher)

| Net income | Low income | Medium income lower | High income |
|------------|------------|---------------------|-------------|
| 26.64      | -0.4276†   | -0.4694†            | -0.4565§    |
|            | (0.1004)   | (0.0999)            | (0.1188)    |
| 14.90      | -0.161     | -0.1743             | -0.2504§    |
|            | (0.1080)   | (0.1077)            | (0.1270)    |
| 21.19      | -0.4551†   | -0.4318†            | -0.5646†    |
|            | (0.0945)   | (0.0940)            | (0.1083)    |
|            |            |                     | (0.2024)    |

### Marital status (ref. Married, No information)

| Marital status | Divorced | Single | Widowed |
|----------------|----------|--------|---------|
| 2.99           | 1.9632†  | 1.492† | 1.1739† |
|                | (0.1289) | (0.0883) | (0.2184) |
| 16.90          | 1.841†   | 1.461† | 1.0208† |
|                | (0.1259) | (0.0873) | (0.2156) |
| 23.39          | 2.3574†  | 1.685  | 1.512†  |
|                | (0.1599) | (0.1041) | (0.2843) |

### Car engine capacity, litres (ref. (1.0-1.4])

| Car engine capacity, litres | (1.4-1.6] | 1.6+ |
|-----------------------------|-----------|-----|
| 10.37                       | -0.1263   | 0.5933† |
|                             | (0.1541)  | (0.1018) |
| 16.00                       | 0.0925    | 0.5933† |
|                             | (0.1226)  | (0.1016) |
| Missing information         | 0.5933†   | 0.5933† |
|                             | (0.1018)  | (0.1016) |

### Phone provided (ref. Work and home number provided)

| Phone provided | Work number provided | Home number provided | No phone number provided |
|----------------|----------------------|----------------------|-------------------------|
| 13.35          | 0.1019               | 0.4469†             | 0.3299†                 |
|                | (0.1073)             | (0.0847)            | (0.1114)                |
| 31.20          | 0.1148               | 0.4762†             | 0.3482†                 |
|                | (0.1069)             | (0.0846)            | (0.1109)                |
| 15.12          | 0.1637               | 0.4955†             | 0.3487§                 |
|                | (0.1218)             | (0.0984)            | (0.1283)                |
|                | -0.1986              | 0.351§              | 0.2701                  |
|                | (0.2447)             | (0.1713)            | (0.2351)                |

### Profession/Occupation (ref. Gender neutral)

| Profession/Occupation | Female profession | Male profession |
|-----------------------|-------------------|-----------------|
| 5.89                  | -0.5111§          | -0.2709§        |
|                       | (0.1938)          | (0.1134)        |
| 13.08                 | -0.6108§          | -0.124§         |
|                       | (0.1928)          | (0.1129)        |
|                       | -0.7843§          | -0.2832§        |
|                       | (0.2827)          | (0.1246)        |
|                       | -0.2068           | -0.2767         |
|                       | (0.2653)          | (0.3003)        |

### Models' Fit Statistics

| Models' Fit Statistics | AIC | Cox&Snell Pseudo-R² | Nagelkerke Pseudo-R² |
|------------------------|-----|---------------------|----------------------|
| AIC                    | 10838.202 | 0.0600              | 0.3823               |
| Intercept and Covariates | 6976.242 | 0.0600              | 0.3823               |
| 8467.386               | 7003.602 | 0.0595              | 0.3796               |
| 2551.084              | 5117.254 | 0.0707              | 0.4253               |
| 1833.609              | 1833.609 | 0.0337              | 0.2606               |
The variables that show significant statistical effects are consistent with the general literature on credit scoring. Shorter loan duration and longer time in employment are associated with the lower PD (Hand, 1998; Thomas et al., 2017). Females as well as customers having a job typical for women are more creditworthy, similarly to married applicants, and those that provide both commercial and home phone numbers. Having kids increases the chances of default.

Considering the features of the car purchased, customers buying cheaper vehicles, with medium engine, having the down payment over 50% of the car price, are less likely to default. Whereas in case of the customers buying older cars, the risk is rising.

When Gender is removed (Model without Gender), small changes in the parameter estimates can be observed, e.g. the parameter estimates go down for Marital Status and Net Income, but increase for Loan Duration and Time at Employment. This illustrates the situation of omitted variable bias when the protected characteristic is not included in the model, yet it is associated with the remaining predictors and the dependent variable. The parameter estimates in Model without Gender (Model 2) partially reflect the effect of Gender.

In order to understand whether there are differences in the risk profiles of the two sexes, separate modes have been fitted to men and women (Column 5 and 6 in Table 2 – Model 3 for men and Model 4 for women). Whilst there are some changes in the estimated parameters between Model 3 for men only and Models 1&2 (with and without Gender), the variables remain the same, except for some differences in the magnitude of the parameter estimates. In Model 4 for women only 7 of 25 categories significant in Model 1 become insignificant in Model 4, which may be the result of being a small segment in the sample. On the contrary, 3+ kids, which is not significant in the Model 3 for men only, becomes highly significant in Model 4 (female). For other categories that remain significant, there is a dramatic change in magnitude, e.g. the effect of Single in Marital Status for females is almost half of that for males. This indicates the difference with the Model 3 for men and Models 1&2 (with and without Gender) that are dominated by a larger male segment (around 75%).

Measures of model fit include Akaike Information Criterion (AIC) with lower values indicating better fit, and pseudo-R2. Models 1 & 2 (with and without Gender) have been estimated on the same sample, therefore model fit measures can be compared directly, and Model 1 with Gender demonstrates better fit. This, however, does not translate into superior predictive accuracy (Table 3), where Model 2 without Gender is slightly better, but the difference is an artefact of random
variation (as will be shown in Section 4.3). Models 3 & 4 for men and women have been estimated on different segments, therefore are not directly comparable, yet judgment can be made relative to the corresponding measures for models with intercept and intercept + covariates.

Table 3. Measures of predictive accuracy for the training and test sample: Area under the ROC curve (AUC), Sensitivity (proportion of correctly predicted defaults), 1-Specificity (proportion of incorrectly predicted non-defaults). For Sensitivity/1-Specificity the cut-off was selected at the mean score of the corresponding training sample.

|                | Total sample | Male only segment | Female only segment |
|----------------|--------------|-------------------|--------------------|
|                |              | Model 1 with Gender | Model 2 without Gender | Model 3+4 for men and women | Model 1 with Gender | Model 2 without Gender | Model 3 for men | Model 1 with Gender | Model 2 without Gender | Model 4 for women |
| AUC training sample | 0.92066 | 0.92111 | 0.92381 | 0.93341 | 0.93316 | 0.93330 | 0.87300 | 0.87390 | 0.88596 |
| Sensitivity    | 0.85473 | 0.85005 | 0.85848 | 0.89138 | 0.87485 | 0.87603 | 0.71364 | 0.75455 | 0.79091 |
| 1-Specificity  | 0.14502 | 0.15523 | 0.15281 | 0.15938 | 0.14855 | 0.14732 | 0.14314 | 0.17347 | 0.16780 |
| AUC test sample | 0.89014 | 0.88984 | 0.89433 | 0.91465 | 0.91390 | 0.91490 | 0.79651 | 0.79434 | 0.80615 |
| Sensitivity    | 0.79041 | 0.78277 | 0.79026 | 0.83962 | 0.82076 | 0.83019 | 0.61818 | 0.63636 | 0.65455 |
| 1-Specificity  | 0.15298 | 0.15432 | 0.15112 | 0.16378 | 0.15126 | 0.15100 | 0.12351 | 0.16269 | 0.15504 |

For Area under the curve (AUC) reported in Table 3 comparisons should also be made within the same sample/segment. For the total sample the column ‘Model 3+4 for men and women’ refers to estimated PDs obtained from segmented models, but then combined into a single score to make it comparable to AUC for Models with and without Gender (Models 1 and 2). Segmented PDs combined do show an uplift in predictive performance, albeit a modest one. When looking at predictive accuracy of models applied separately to men or women, for men there is practically no difference between the three PDs, i.e. from Model 1, Model 2 and Model 3. The greatest difference is observed for women. Whilst there is little difference in predictive accuracy for women from adding Gender into Model 1, the segmentation does allow capturing unique features of female risk profiles, hence a modest uplift is observed. The formal tests of difference between AUCs will be reported in Section 4.3.

Yet the ultimate question is what this means for the chances to be accepted for credit. Hand (2012b) outlines the following scenarios when talking of potential solutions in achieving discrimination-free credit decisions:

a. **Current situation** – prohibit the use of Gender. One can also consider removing variables with Gender, but the question arises, what level of correlation would be acceptable and how many variables would be left for model building.
b. **Ensure equal outcome** – accept the same proportion of men and women, which can be easily achieved with separate Models 3 and 4 for men and women.

c. **Eliminate the effect of gender statistically** by estimating the regression model with *Gender*, but removing its coefficient when applying to new applicants. In this way, because of statistical control, the remaining coefficients in the model will indicate the effect on the default given the effect of *Gender*.

Note that only scenario A is legal under the existing regulations, since B and C require the use of *Gender* in model-building, and therefore, we focus on it in other further investigation.

To assess the impact on access to credit, we apply Models 1 and 2 (with and without *Gender*) and calculate proportions of men and women rejected by each of the two models for different cut-off levels that would correspond to a range of rejection/acceptance rates: from 0.1 to 0.9 in 0.1 increments (Figure 1), although it should be noted that rejection rates above 0.6 or 60% are, perhaps, not realistic. In this part of the analysis we address the concern raised by two anonymous reviewers who quite rightly noted that in families it is usually men who apply for credit, so it is unfair to claim that all female applicants are disadvantaged, since for many married women, their husbands will apply for credit on their behalf. Therefore, we look at rejection rates for unmarried customers and compare rejected proportions for men and women when they are scored either with Model 1 or Model 2. Unmarried customers include single, divorced and widowed, and constitute 21.57% from total (with 8.91% of women and 12.66% of men from total).

E.g., if a lender rejects 60% of the total sample (shown on the vertical axis and the top row under the horizontal axis) and uses PDs from Model 2 without *Gender* as scores, 65% of all men in the sample would be rejected as compared to 54% of all women (shown on horizontal axis, rows 1 and 2 respectively in the data table). However, if Model 1 with *Gender* is used for the same cut-off (60% overall rejection), the corresponding percentages become 71% for men and 45% for women, thus increasing chances of women to be accepted for credit and rewarding them for being better credit risks.

Overall, men, being less creditworthy, benefit from Model without *Gender* (Model 2). On the contrary, women would benefit from including *Gender*, since more females would be accepted for credit. However, it is important to note the removal of *Gender* does not make the reject rates equal for both sexes.
Figure 1. Impact on rejection by gender when different models are used, considering different proportions of men/women rejected versus the overall reject rate, unmarried customers. Vertical axis and the top row on the horizontal axis show overall reject proportion; values in the data table show reject proportions for a given group and model.

4.2. Inter-relations of Gender and other predictors in the models

The results in the previous section demonstrate that although Gender has a statistically significant effect on the PD, its removal from the model does not change the overall predictive accuracy as measured by AUC (although there are changes in the rejection rates for unmarried women). The apparent lack of change in the predictive accuracy can be explained by the association of Gender with other predictors in the model, so when it is removed, other variables that remain in the model, act as a proxy for Gender. In this section we explore the association between predictors in order to understand how this happens.

We use two different techniques: Bayesian networks (BN); Multiple correspondence analysis (MCA).
4.2.1. Bayesian networks

Bayesian networks (BN) are probabilistic models represented by graphs, where random variables are nodes and conditional dependencies between them are arrows/ arcs. The graph separates the joint (or global) probability distribution of the set of nodes \( V = \{X_1, \ldots, X_v\} \), into a set of local probability distributions, one for each variable. This relies on the Markov property of BNs (Korb and Nicholson, 2010), which implies that a random variable \( X_i \) directly depends only on its parents \( pa(X_i) \). In our case we have multinomial (categorised) data, therefore, global and local distributions are given as probability or contingency tables, and

\[
P(X_1, \ldots, X_v) = \prod_{i=1}^{v} P(X_i | pa(X_i))
\]

The above formula assumes conditional independence. BNs are often used as causal models even when estimated from observational data (Pearl, 1988). Specifying a causal model of credit default is beyond the scope of this paper. In this section the objective is to identify associations between Gender and other variables in the models, and it is sufficient to learn the skeleton of the network or to estimate the undirected essential graph underpinning the network structure. This is the first step in constrained-based algorithms that conduct conditional independence tests in order to establish relations defined by the Markov property above (Scutari, 2010).

Specifically, in order to understand the structure of connections in our training sample, we use Max-Min Parents and Children (MMPC) algorithm (Tsamardinos et al., 2006) as implemented in R package bnlearn (R Development Core Team, 2010; Scutari, 2010). It considers all possible pairs of parents-children and then removes those that are not directly connected. It uses the mutual information between categorical variables (Tsamardinos et al., 2006, R Development Core Team, 2010). The rest of the analysis in this paper has been done using SAS software.

Figures 2-4 below show Markov blankets (collection of the node’s direct connections) for the variables of interest. Figure 2 presents the Markov blanket for Flag, the default indicator, and the dependent variable that is of the main interest in credit risk modelling. Figure 3 shows the Markov blanket for Gender, which is of the main interest for this study. Figures 2 and 3 demonstrate that Gender is not directly connected to Flag. But Flag is directly related to Down payment, Car Age and Marital Status. Marital Status in its turn is directly related to Gender. Gender is also directly linked to Occupation and Net Income. When Gender is removed from the model (Figure 4), the
direct link between Flag and Car Age only remains. Therefore, even if Gender is not directly related to the outcome, its removal affects the structure of associations between the remaining variables.

Figure 2. Markov blanket for Flag, default indicator (Model 1). Red lines show direct connections. Variables with underlined names in red are directly related to Flag. Full variables names: occup – Occupation/Profession; phone - Phone number provided; engine_c - Car engine capacity; marital – Marital status; net_inc – Net Income; toj - Time in employment; term – Loan duration; carage – Car age; downp – Down payment; price_c – Car price; child – No of children.

Electronic copy available at: https://ssrn.com/abstract=3212702
Figure 3. Markov blanket for Gender.

Figure 4. Markov blanket for Flag (default indicator) with Gender removed (Model 2).
4.2.2. Multiple Correspondence Analysis

Based on the dependence structure revealed in the previous section, this section explores the relationship between Gender and other predictors with Multiple Correspondence Analysis (MCA). In order to make results easier to interpret and present graphically, we have restricted the variables used in this section to those that are directly related to Gender in Markov blanket (Figure 3), because among all 12 variables used in Gender’s Markov Blanket only some of them are directly related to Gender. Besides, it is very difficult to provide a concise interpretation when the analysis is done on all 13 variables with the corresponding categories. In addition, No. of children has also been included given its importance in describing the family status of credit applicants, and therefore, relevance for this study. No. of children is directly related to Marital Status and Occupation, which are in Markov blanket of Gender.

MCA allows analysis of the relationship of multiple categorical variables similar to Principal Component Analysis (PCA) for numeric variables (Beh and Lombardo, 2014; Greenacre, 2016; Sourial et.al, 2010). The result of the MCA is a graphical representation of a contingency table in order to show associations between the qualitative variables of interest in a low-dimensional space. This shows the patterns which cannot be revealed with pairwise analysis. The association or proximity is based on chi-square statistic or inertia, which is decomposed into eigenvalues, similar to how the variance is decomposed in PCA.

The chi-square statistic summarises the deviations in a contingency table between the observed frequencies and the frequencies expected under the assumption of independence /homogeneity of the categorical variables:

\[ \chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \],

where \( n_{ij} \) is observed value in row \( r \), and column \( c \); and \( e_{ij} \) is the corresponding expected value.

The inertia is a measure of deviation from homogeneity (or how much variation is contained in the table), which is not dependent on the sample size \( n \):

\[ \phi^2 = \frac{\chi^2}{n} \].

Following Greenacre (2016), let \( N \) be a \( I \times J \) non-negative data matrix. The data matrix is converted to the correspondence matrix \( P \), which is a matrix of relative frequencies:

\[ P = \frac{1}{n} N \],

where \( n = \sum_i \sum_j n_{ij} = 1^T N 1 \).
Row and column masses (or row and column marginal proportions) are given by
\[ r = P_1 \quad \text{and} \quad c = P^T 1 \]  
and diagonal matrices of row and column masses:
\[ D_r = \text{diag}(r) \quad \text{and} \quad D_c = \text{diag}(c) \]  
Row profiles are contained in \( R = D_r^{-1} P \),  
where the elements of each row sum to one. Each \((i,j)\) element of \( R \) is the observed probability of being in column \( j \) provided it is in row \( i \).

Column profiles are contained in \( C = D_c^{-1} P \).  
The correspondence analysis (CA) algorithm computes the coordinates based on generalised singular value decomposition of \( P \)
\[ P = A D_a B^T \]  
where \( A^T D_r^{-1} A = B^T D_c^{-1} B = I \)  
and \( D_a \) is the diagonal matrix of singular values, so that \( \alpha_1 > \alpha_2 > \alpha_3 > \ldots \)
The principal inertias are given by
\[ \lambda_k = \alpha_k^2, \quad k = 1, 2, \ldots, K \]  
where \( K = \min \{I-1, J-1\} \).

In MCA
\[ P = B D_a^2 B^T \]  
because in this paper we use the version of MCA based on the Burt matrix, which gives all two-way cross-tabulations of the \( Q \) categorical variables with a number of categories \( J = \sum q_j \).

For more details on the theory of CA and MCA, please, refer to Greenacre (2016).

Table 4 shows decomposition of total inertia with corresponding statistics, with the singular value indicating the relative contribution of each dimension to an explanation of the inertia, or proportion of variation. The principal inertia is an indicator of how much of the variation in the original data is retained in the dimensional solution. Table 4 indicates that all variation (or 100% of total inertia) in Burt matrix can be summarised by 13 dimensions. The percent of each dimension in total inertia (as shown in column ‘Percent’ of Table 4) ranges as shown in from approximately 11 to 5, with no dimension encompassing the major part of the data, which should alleviate concerns about excessive multicollinearity. Nevertheless, there are some high contributions of categories into selected dimensions as shown in Table 5, which presents three of the largest dimensions (detailed results on other dimensions are available on request). Here the main variable of interest Gender only shows substantial loading to the first two dimensions: female is positively
associated with Dimensions 1 and 2, with male showing negative association, therefore, here the differences between sexes are most apparent. Dimension 3 shows high loadings/associations of Net Income.

Table 4. Inertia and chi-square decomposition

| Singular Value | Principal Inertia | Chi-Square | Percent | Cumulative Percent |
|----------------|-------------------|------------|---------|--------------------|
| 0.5533         | 0.2844            | 93455      | 10.94   | 10.94              |
| 0.4923         | 0.2424            | 79651      | 9.32    | 20.26              |
| 0.4655         | 0.2167            | 71226      | 8.34    | 28.60              |
| 0.4594         | 0.2111            | 69369      | 8.12    | 36.72              |
| 0.4530         | 0.2052            | 67436      | 7.89    | 44.61              |
| 0.4493         | 0.2019            | 66355      | 7.77    | 52.38              |
| 0.4467         | 0.1995            | 65570      | 7.67    | 60.05              |
| 0.4456         | 0.1986            | 65260      | 7.64    | 67.69              |
| 0.4434         | 0.1966            | 64595      | 7.56    | 75.25              |
| 0.4385         | 0.1923            | 63201      | 7.40    | 82.65              |
| 0.4218         | 0.1779            | 58464      | 6.84    | 89.49              |
| 0.3846         | 0.1479            | 48614      | 5.69    | 95.18              |
| 0.3543         | 0.1255            | 41259      | 4.82    | 100.00             |
|                | 2.6000            | 854455     | 100.00  |                     |

Table 5. Contribution of each category to each dimension

| Variable     | Category                                      | Dim1     | Dim2     | Dim3     |
|--------------|-----------------------------------------------|----------|----------|----------|
| Gender       | female                                        | 0.8232   | 0.9901   | -0.0875  |
|              | male                                          | -0.3001  | -0.3609  | 0.0319   |
| No of children| 1 kid                                         | -0.6188  | 0.7525   | -0.2837  |
|              | 2 kids                                        | -0.971   | 0.3206   | -0.306   |
|              | 3+ kids (3 and more children)                 | -1.2519  | -0.1746  | 0.5808   |
|              | mis kids (missing information about number of children) | 0.1841 | -0.1729 | 1.8177 |
|              | no kids (no children)                         | 0.6476   | -0.4171  | -0.2174  |
| Occupation   | female occ (occupations where majority are women) | 0.8005 | 1.6965   | 0.6333   |
|              | male occ (occupations where majority are men) | 0.0249   | -1.0634  | 0.1797   |
|              | neut occ (gender neutral occupations)         | -0.0622  | 0.0484   | -0.075   |
| Marital status| D (divorced)                                 | 1.084    | 2.3252   | 1.4982   |
|              | M (married)                                   | -0.393   | 0.0406   | -0.1278  |
|              | S (single)                                    | 1.4648   | -0.789   | 0.3786   |
|              | W (widowed)                                   | 1.532    | 1.2036   | -0.2901  |
| Income       | Low inc (low income)                          | 0.3809   | -0.0413  | -1.0002  |
|              | Mid inc1 (lower middle income)                | 0.1369   | 0.0816   | -0.4144  |
|              | Mid inc2 (higher middle income)               | -0.0633  | -0.0327  | 0.3604   |
|              | High inc (high income)                        | -0.4638  | 0.0519   | 0.9153   |
Interpretation of the results is complemented by Figures 5-7 where the judgment depends on how the variables’ categories are relatively close to each other. In Figure 5, the first dimension reflects a split of the clients on Marital Status and No of Children. Two categories of Marital Status: Divorced (D) and Widowed (W) are located on the top of Figure 5, and two others categories: Married (M) in the middle and finally the category Separated (S) much below. Where in case of No of Children, the three categories: 1, 2 or 3+ kids are far away from No kids and Missing information (these two are close to each other). Dimension 2 (Figure 7) separates Gender categories, putting them on opposite sites, namely female on the right and male on the left. Furthermore, Dimension 2 separates Female occupation (top right corner on) from two remaining categories: Neutral and Male occupation. Dimension 3 (Figures 6 and 7) clearly reflects a division of the income categories, namely Low (Low inc) and Lower middle income (Mid inc1) versus two other categories, that is: Higher middle income (Mid inc2) and High income (High inc).

Variable categories which are related in some way, appear in the same quadrants, although the distance is approximate. For example, the top right quadrant of the Figure 5 shows that the following variable categories are associated: Female, Female occupation, Widowed and Divorced. Similarly, in the bottom left quadrant Male and 3+ kids appear together. The other categories close to Male, e.g. Higher middle income (Mid inc2) or Neutral occupation (neut occ) are almost at the origin, so there is little association. In Figure 7 Male appears on the origin of Dimension 3 (the one that is associated with Net Income), Female is not far from the origin, but still appears on the same side with Low income (Low inc) and Lower middle income (Mid inc1). One can say that female borrowers tend to be widowed or divorced and with lower income as compared to males. This makes the impact of Gender removal from risk assessment even more socially significant, it is not simply the females in general who are disadvantaged by this policy, there are socially vulnerable groups that may have restricted access to credit.
Figure 5. MCA analysis, position of categories against Dimension 1 and Dimension 2. The distance is approximate, represents relative position of categories from the origin. Full category names are given in Table 5.
Figure 6. MCA analysis, Dimension 1 and Dimension 3. The distance is approximate, represents relative position of categories from the origin. Full category names are given in Table 5.

Figure 7. MCA analysis, Dimension 2 and Dimension 3. The distance is approximate, represents relative position of categories from the origin. Full category names are given in Table 5.
4.3 Comparison of AUC

In this section we follow another suggestion from one of the anonymous referees and investigate if the predictive accuracy changed when variables correlated with Gender are removed from the model. In fact, such situation may arise if regulators require not only prohibited variables to be removed from the statistical models, but also variables correlated with the prohibited ones. A possible choice is to use variables in the Markov blanket for Gender, i.e. those directly connected to it, as identified is Section 4.2.1: Marital Status, Net Income, Occupation. We label these as Models 1a/2a. We also subjectively add to the list of removed variables No of children because of its connection with Marital Status, and label them 1b/2b. The full list of Models developed and tested is given below:

Model 1 - all variables, including Gender
Model 2 – no Gender
Model 1a – no Marital Status, Net Income, Occupation
Model 2a – no Gender, Marital Status, Net Income, Occupation
Model 1b – no Marital Status, Net Income, Occupation, No of children
Model 2b – no Gender, Marital Status, Net Income, Occupation, No of children
Model 3+4 – combined PDs from Model 3 (male) and Model 4 (female), see Section 4.2.

The receiver operating characteristic curves we are going to compare are developed and applied to the same train and test samples, so they and their corresponding AUCs are correlated. Therefore, we used the test for correlated curves proposed by DeLong et al. (1988), which tests that the selected pair of AUCs are significantly different against the null hypothesis of no difference.

As noted in Section 4.2, the removal of Gender (Model 1 v Model 2) does not affect the predictive accuracy and this is confirmed by high p-values > 0.05. However, Model 1 significantly differs from all other models, although it should be noted that the difference with segmented model (3+4) becomes insignificant on the test sample. So the removal of variables correlated with Gender affects the predictive accuracy of credit scoring models. If Gender is then added to the model with correlated variables removed (Model 1a v Model 2a), the result is insignificant on the test sample. The same applies to modes without No of children (Model 2a v 2b). Therefore, we cannot conclude that Gender is a powerful predictor, despite it being a statistically significant variable in Model 1.
This one of the questions that can be explored further: there might be weak associations with other variables that become significant in aggregate, or credit scoring models may not be sensitive to removal of just one variable (although it may depend on the exact combination of variables in the model).

Table 6. Results of the tests of difference in predictive accuracy of modes with different variables. Values in the middle are p-values for pairwise test of difference against the null hypothesis of no difference: Pr > ChiSq test statistic.

|          | AUC | Model 1 | Model 2 | Model 1a | Model 2a | Model 1b | Model 2b | Model 3+4 |
|----------|-----|---------|---------|----------|----------|----------|----------|-----------|
|          |     |         |         |          |          |          |          |           |
| **Train** |     |         |         |          |          |          |          |           |
| Model 1  | 0.9207 | -       | 0.3342  | <.0001   | <.0001   | <.0001   | <.0001   | 0.0041    |
| Model 2  | 0.9211 | 0.3342  | -       | <.0001   | <.0001   | <.0001   | <.0001   | 0.0014    |
| Model 1a | 0.9117 | <.0001  | <.0001  | -        | 0.0391   | 0.4946   | 0.1763   | <.0001    |
| Model 2a | 0.9119 | <.0001  | <.0001  | 0.0391   | -        | 0.0273   | 0.3525   | <.0001    |
| Model 1b | 0.9104 | <.0001  | <.0001  | 0.4946   | 0.0273   | -        | 0.0599   | <.0001    |
| Model 2b | 0.9107 | <.0001  | <.0001  | 0.1763   | 0.3525   | 0.0599   | -        | <.0001    |
| Model 3+4 | 0.9238 | 0.0041  | 0.0014  | <.0001   | <.0001   | <.0001   | <.0001   | -         |

|          |     |         |         |          |          |          |          |           |
| **Test** |     |         |         |          |          |          |          |           |
| Model 1  | 0.8901 | -       | 0.7704  | 0.0074   | 0.0011   | 0.0093   | 0.0016   | 0.2034    |
| Model 2  | 0.8898 | 0.7704  | -       | 0.0127   | 0.0020   | 0.0103   | 0.0015   | 0.1822    |
| Model 1a | 0.8809 | 0.0074  | 0.0127  | -        | 0.1150   | 0.6428   | 0.2517   | 0.1632    |
| Model 2a | 0.8813 | 0.0011  | 0.0020  | 0.1150   | -        | 0.0994   | 0.6378   | 0.0464    |
| Model 1b | 0.8785 | 0.0093  | 0.0103  | 0.6428   | 0.0994   | -        | 0.1255   | 0.1521    |
| Model 2b | 0.8789 | 0.0016  | 0.0015  | 0.2517   | 0.6378   | 0.1255   | -        | 0.0392    |
| Model 3+4 | 0.8943 | 0.2034  | 0.1822  | 0.1632   | 0.0464   | 0.1521   | 0.0392   | -         |

5. Concluding remarks

This paper has explored the concept of “equal treatment”, which is the fundamental principle of anti-discrimination regulations, in application to statistical or algorithmic discrimination in automated decision-making. We have used the case of retail credit risk screening and examined “equal treatment” of men and women, which translates into the prohibition to use Gender in credit scoring models. The potential impact of the prohibition relates to lenders (any effects on predictive power of models, or their ability to distinguish between Good/Bad borrowers), and consumers (their chances to be accepted/rejected for loans). Our illustration of potential impact is based on a dataset of applications for auto loans, where women form a minority and constitute about a quarter of the portfolio. Following a standard credit scoring methodology, four models have been built initially: Model 1 with and Model 2 without Gender as a dummy variable, and two separate models for men and women (Model 3 and 4 respectively).


*Gender* is statistically significant as a dummy variable, yet its removal does not have a pronounced effect on discriminatory power of the model. This is due to the correlation/association of *Gender* with other characteristics that are not legally restricted and remain in the model. If predictive accuracy is unaffected, lenders are able to maintain similar levels of bad debt for a given acceptance level irrespective of which model is used, therefore, there is no impact on lenders. It does not mean though that they will accept the same applicants under different models.

The paper then concentrates on unmarried applicants to investigate the impact for consumers and shows that female applicants, being more creditworthy, benefit from Model 1 with *Gender*, which gives them extra points for being *Good* risks and therefore, increases their chances to be accepted for credit. Their rejection rates are lower compared to those for men. When applying unisex Model 2 (without *Gender*), the chances to be accepted for credit decrease for women but increase for men, although in general, females still exhibit lower rejection rates as compared to men. So the prohibition does not lead to the equality in the outcome. It should be noted though that if the desire is to reject the same % of men and women, this can be easily achieved when both groups are treated separately (Models 3 and 4) and the same proportions of both sexes that show better ratings on the corresponding scores are selected for credit.

The results are indicative of the law of unintended consequences. Surely, the main objective of the equality provisions is to protect disadvantaged groups of consumers. Yet, it has been shown that the regulations do not ensure equality of outcome. More creditworthy groups subsidise worse risks, and it could be questioned if this subsidy is justified.

Another potential pitfall for achieving equality are situations, when protected groups constitute a minority, and therefore, are not equally represented in the data, and one can argue, not equally treated, as the result of this. In our case the unisex Model 2 is dominated by a majority group, and minority unique features may not be captured by a ‘politically correct’ model.

One more controversy that we would like to highlight is the difficulty of removing the effect of prohibited variables. We have shown that *Gender* is significantly associated with other ‘legal’ variables, therefore, even when removed, the prohibited variable still partially influences the model through associations. This could explain the relative insensitivity of predictive accuracy to *Gender* removal and the remaining difference in reject rates between men and women. However, there is no clear legal guidance on potential solutions in such situations, i.e. what level of correlation would be (un)acceptable?
In general, the paper has highlighted certain inconsistencies in the existing framework when it is being applied to statistical discrimination. These highlights are particularly timely given the increasing applications of machine-learning algorithms in decision-making. Nevertheless, this study is not without limitations. An obvious one is we investigate the accept/reject decision by using a portfolio of accepted loans. This is, however, a well-known problem in credit scoring: the models are developed on accepted loans, since for rejected ones the outcome variable is not observed. There is a methodology to correct for a potential sample selection bias called ‘reject inference’, but Crook and Banasik (2004) have investigated a variety of such approaches on a rarely available unbiased population and concluded that reject inference was not effective. Therefore, we believe, that our investigation of predictive accuracy is still valid. As for the impact on the consumers, further research could investigate the whole customer populations, should rejected applications become available. Yet accepted customers form part of this population, so one can argue our results hold for at least part of the overall population.

Another limitation consists in the fact that the paper has explored only a limited number of variables. Although these are typical variables as stated by the credit scoring literature (Hand, 1998, Thomas et al., 2017), they are by no means exhaustive in terms of other variables that can be used in different portfolios, countries and credit products. Our study can be viewed as an exploratory illustration to highlight the issues that would warrant further research.

Further investigation of other protected characteristics across a wider range of portfolios and countries would contribute towards finding better understanding of discrimination mechanism and fairer solutions. Other areas of decision-making, such as fraud detection, can also be investigated.
References

Agarwal, S., He, J., Sing, T. F. and Zhang, J. (2016) Gender Gap in Personal Bankruptcy Risks: Empirical Evidence from Singapore. *Review of Finance*, rfw063. (Available from https://doi.org/10.1093/rof/rfw063).

Anderson, R. (2017) The Credit Scoring Toolkit. Oxford University Press.

Andreeva, G., Ansell, J. and Crook, J.N. (2004) Impact of Anti-Discrimination Laws on Credit Scoring, *Journal of Financial Services Marketing*, 9, 22-33.

ATI (2017) Written evidence submitted by The Alan Turing Institute (ALG0073) to ‘Algorithms in decision-making’ inquiry. (Available from http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69165.html).

Baesens B., Mues C., Martens D. and Vanthienen, J. (2009) 50 years of data mining and OR: upcoming trends and challenges. *The Journal of the Operational Research Society*, 60, SI6-S23.

Banasik, J.L., Crook, J.N., and Thomas, L.C. (1996) Does scoring subpopulations make a difference? *International Review of Retail, Distribution and Consumer Research*, 6, no. 2, 180 - 195.

Becker, G.S. (1971) The Economics of Discrimination. University of Chicago Press.

Beh, E.J. and Lombardo, R. (2014) Correspondence Analysis: Theory, Practice and New Strategies. John Wiley & Sons.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M.J., Morgenstern, J., Neel, S. and Roth, A. (2017) A Convex Framework for Fair Regression. *ArXiv*, https://arxiv.org/abs/1706.02409

Bijak, K., and Thomas, L. C. (2012) Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 39, no. 3, 2433-2442. DOI: 10.1016/j.eswa.2011.08.093

Chan, W.L. and Seow, H. (2013) Legally Scored. *Journal of Financial Regulation and Compliance*, 21, 39 – 50.

Charles, K.K. and Guryan, J. (2011) Studying Discrimination: Fundamental challenges and recent progress, *NBER Working paper*, 17156. (Available from: http://www.nber.org/papers/w17156).

Coin, D. (2013) Are female entrepreneurs better payers than men?, *Questioni di Economia e Finanza, Banca D’Italia*, Number 186, June 2013. (Available from: http://www.bancaditalia.it/pubblicazioni/econo/quest_ecofin_2/qef186/QEF_186.pdf).
Crook, J. and Banasik, J. (2004) Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance, 128*, 857–874.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1998) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*, no. 3, 837-45.

D’Espallier, B., Guérin, I. and Mersland, R. (2011) Women and Repayment in Microfinance. *World Development, 39*, 758–772.

Do, C. and Paley, I. (2013) Does Gender Affect Mortgage Choice? Evidence from the US. *Feminist Economics, 19*, 22–68.

ECOA (1976) Regulation B - Equal Credit Opportunity, *15 US Code, §§ 1691-1691f*, (Available from http://www.ftc.gov/ogc/stat3.htm).

European Union Council (2004) COUNCIL DIRECTIVE 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, (Available from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:en:PDF).

European Parliament, (2011) The use of gender in insurance pricing, Directorate General for Internal Policies, (Available from http://www.europarl.europa.eu/RegData/etudes/note/join/2011/453214/IPOL-FEMM_NT(2011)453214_EN.pdf).

Fair, W. (1979) Hearings before the Subcommittee on Consumer Affairs of the Committee on Banking, Housing and Urban Affairs, United States Senate, 96th Congress, 1st Session on S. 15, June 4 and 5, 1979. US Government Printing Office: Washington, DC.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2017) Predictably Unequal? The Effects of Machine Learning on Credit Markets. SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038

Greenacre, M., (2016) Correspondence Analysis in Practice. Interdisciplinary Statistics Series, third Edition, Barcelona: Chapman&Hall/CRC.

Hand, D. J., Henley, W. E. (1997) Statistical classification methods in consumer credit scoring: a review. *J. R. Statist. Soc. A, 160*, 523–541.

Hand, D.J. (1998) Consumer credit and statistics, in *Statistics in Finance*, ed. Hand, D. J. and Jacka, S. D., Arnold, E., London: Arnold.

Hand, D. J. (2009) Modern statistics: the myth and the magic (RSS Presidential Address), *J. R. Statist. Soc. A, 172*, 287-306.
Hand, D.J. (2012a) Credit scoring, insurance and discrimination. In Statistics, Science, and Public Policy XVI: Risks, Rights, and Regulations. April 17-20, 2011, ed. A.M. Herzberg, Queen's University, Canada, 85-90.

Hand, D.J. (2012b) Confusion in scorecard construction – the wrong scores for the right reasons. Presentation at Model Risk in Retail Credit Scoring – Statistical Issues. London. (Available from: http://wwwf.imperial.ac.uk/~abellott/ModelRiskWorkshop.html).

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29 - 36.

HMDA (2015) Home Mortgage Disclosure (Regulation C), 66128 Federal Register / Vol. 80, No. 208 / Wednesday, October 28, 2015 / Rules and Regulations, https://www.gpo.gov/fdsys/pkg/FR-2015-10-28/pdf/2015-26607.pdf

Johnson, R.W. (2004) Legal, social and economic issues in implementing scoring in the United States. In: Thomas LC, Edelman DB and Cook JN (eds.) Readings in Credit Scoring. Oxford: Oxford University Press.

Jung, K.M. and Thomas, L.C. (2008) A Note on Coarse Classifying in Acceptance Scorecards. The Journal of the Operational Research Society, 59, 714-718

Kohavi, R. (1995) A study of cross-validation and boot-strap for accuracy estimation and model selection. Proceedings of International Joint Conference on Artificial Intelligence.

Korb, K., and Nicholson, A.E. (2010) Bayesian Artificial Intelligence. Chapman & Hall/CRC, 2nd edition.

Lin, S-M., Ansell, J. and Andreeva, G. (2012) Predicting default of a small business using different definitions of financial distress. Journal of the Operational Research Society, 63, 539-548

McDonald, S. (2015) Indirect Gender Discrimination and the ‘Test-Achats Ruling’: An Examination of the UK Motor Insurance Market, Presentation at the Royal Economic Society conference 2015, University of Manchester, UK. (Available from: https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RES2015&paper_id=791).

Oxera (2010) The use of Gender in insurance pricing. Analysing the impact of a potential ban on the use of gender as a rating factor. ABI Research Paper, 24, (Available from http://www.oxera.com/Oxera/media/Oxera/The-use-of-gender-in-insurance-pricing.pdf?ext=.pdf).

Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

Phelps, E.S. (1972) The Statistical Theory of Racism and Sexism. American Economic Review, 62, 659 - 661.
R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (Available from http://www.R-project.org/).

RSS (2017) Written evidence submitted by The Royal Statistical Society (ALG0071) to ‘Algorithms in decision-making’ inquiry. (Available from: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69144.html).

Schreiner, M. (2002) Scoring: the next breakthrough in micro-credit, Microfinance Risk Management, Centre for Social Development, Washington University: St. Louis, MO.

Science and Technology Committee (2017) Algorithms in decision-making inquiry. (Available from: https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/inquiry9/).

Scutari, M. (2010) Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35, 1-22.

Siddiqi, N. (2006) Credit risk scorecards: developing and implementing intelligent credit scoring. Wiley: New Jersey.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunanathan, S., Bandeen-Roche, K., Bélard, F. and Bergman, H. (2010) Correspondence analysis is a useful tool to uncover the relationships among categorical variables. Journal of Clinical Epidemiology, 63, 638-646.

Stiglitz, J. E. (1993) Approaches to the Economics of Discrimination, American Economic Review, 62, 287–295.

Taylor, W. (2011) Proving Racial Discrimination and Monitoring Fair Lending Compliance: The Missing Data Problem in Nonmortgage Credit. Review of Banking & Financial Law, 31, 199-264.

Thomas, L.C., Edelman, D.B., and Crook, J.N. (2017) Credit scoring and its applications, 2nd ed. Philadelphia: SIAM Publishing.

Tsamardinos, I., Brown, L.E. and Aliferis, C.F. (2006) The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 65, 31-78.