[SLIDE 2]

Avant de présenter les différentes étapes de l'analyse, nous allons introduire le contexte de cette étude avec quelques chiffres clés sur la filière EdTech en France. La croissance de ce secteur s'est fortement accélérée durant la pandémie, pour atteindre 1.3 milliards de chiffre d'affaires en 2022 et une levée de fonds record de 438 millions d'euros. La même année, le nombre de créations d'entreprises dans ce secteur a été de 34, portant leur nombre total à 430, représentant au total 10 000 salariés. 25% de ces entreprises, nos concurrentes directes donc, sont positionnées exclusivement sur les segments de l'enseignement secondaire et supérieur. La filière EdTech française est relativement concentrée, avec plus de 70% du chiffre d'affaires réalisé par les 20 plus grandes entreprises du secteur ; 85% de ces entreprises se positionnent sur le segment de la formation professionnelle. Le secteur est par ailleurs en voie d'internationalisation, avec 57% des entreprises du secteur déclarant avoir une activité à l'international, qui reste relativement modeste cependant, puisque pour 77% d'entre-elles, cette activité internationale représente moins de 20% du chiffres d'affaires.

[SLIDE 3]

Dans ce contexte, la problématique pour Academy consiste à trouver des débouchés à l'international pour son catalogue de produits existants, qui est constitué de formations en ligne de niveau enseignement secondaire et universitaire, à l'exclusion donc de la formation professionnelle ou des formations diplômantes, sur lesquelles des barrières à l'entrée réglementaires notamment pourraient s'appliquer. Il faudra donc déterminer les pays présentant un bon potentiel à court et moyen terme, et qui seront à prioriser, et ceux présentant un potentiel à plus long terme, pour faire évoluer la stratégie de l'entreprise vers ce que l'on a désigné par « l'objectif 2030 ». Il s'agira donc trouver les indicateurs les plus pertinents pour informer ces décisions, et de les évaluer historiquement et de façon prospective pour déterminer la liste des pays à cibler pour ces différentes phases. Nous avons retenu les thèmes des télécoms et de l'accès internet, du revenu par habitant, de la taille et de la croissance de la population étudiante et du taux de scolarisation secondaire et universitaire comme étant pertinents à notre analyse.

[SLIDE 4]

Afin de répondre à cet objectif, notre analyse s'est articulée autour de 3 étapes : le téléchargement, la vérification et la sélection des données, que nous avons poursuivis par des analyses statistiques et graphiques de nos différents indicateurs, avant de procéder aux recommandations sur les marchés-cibles à prioriser pour les différentes étapes de notre stratégie.

[SLIDE 5]

Le jeu de données fourni, qui était issu de la série EdStatsData de la Banque Mondiale, a nécessité de nombreux retraitements et une exploration détaillée et approfondie pour être utilisable. Ce jeu de données se compose de 5 tables, une table de faits centrale ici, qui contient des données chiffrées pour un certain nombre de pays et d'indicateurs, et 4 tables de dimensions, qui comportent des informations géopolitiques sur les pays, des sources de données par indicateur et par pays, des descriptions détaillées des indicateurs et des notes méthodologiques sur les calculs particuliers effectués pour certaines années. Aucun de ces fichiers de données ne comportait de doublons, mais ils comportaient tous des proportions variables de données manquantes, pouvant aller jusqu'à plus de 86% pour la table des faits. Ces fichiers étaient par ailleurs de grandes dimensions, avec plusieurs centaines de milliers de lignes pour la table des faits et celle des notes méthodologiques, et couvraient une information très diversifiée, avec notamment des données pour la période allant de 1970 à 2100 dans la table des faits, et plus de 3,600 indicateurs regroupés en 37 thèmes distincts dans la table de description des indicateurs. En résumé, nous avions donc à notre disposition un jeu de données de très grandes dimensions, mais à la pertinence discutable et d'une qualité médiocre.

[SLIDE 6]

Afin de rendre ces données, sinon idéales, au moins utilisables pour notre analyse, nous avons effectué un certain nombre de retraitements communs lors de nos opérations d'exploration et de nettoyage, consistant bien évidemment à importer les fichiers et à examiner un échantillon de leur contenu sur les quelques premières lignes,

mais aussi à en décrire les dimensions, à en caractériser la distribution des valeurs manquantes et la présence de doublons, à en déterminer les clés primaires et étrangères en vue de jointures entre les différentes tables et enfin à vérifier et à corriger le type casting des colonnes, et à unifier la nomenclature de nommage de celles-ci.

[SLIDE 7]

Ces différentes opérations nous ont permis, dans la table des pays, d'éliminer les valeurs manquantes, représentées ici par les barres jaunes sur le graphique de gauche, ainsi que les colonnes de données géopolitiques non pertinentes pour notre analyse ; nous avons aussi éliminé un certain nombre de pays, typiquement les pays en guerre ou soumis à des sanctions internationales, ou les lignes correspondant en fait à des groupements de pays, pour ne conserver que 167 pays uniques et 4 colonnes de données descriptives.

[SLIDE 8]

La distribution géographique et économique des pays que nous avons conservés dans notre analyse est représentée sur ces 2 graphiques, avec à gauche, la répartition des pays par niveau de revenus, dominée par les pays à haut revenus qui représentent un peu plus de 32% de notre panel de pays, suivis des pays à revenu moyen supérieur, qui représentent environ 28% de notre échantillon. La treemap à droite représente la répartition géographique des différents groupes de revenus, avec notamment l'Europe et l'Asie centrale dominées par les pays à hauts revenus, et l'Afrique sub-saharienne dominée par les pays à bas revenus.

[SLIDE 9]

La table des sources de données par indicateur et par pays contenait l'intégralité de nos 167 pays et 21 indicateurs les concernant, dont nous avons retenu 4 pour l'analyse, même si, comme on le voit sur le graphique en bas à gauche, la densité informationnelle de ces indicateurs varie grandement, seules les données de

population existant systématiquement pour tous les pays de notre échantillon, les autres indicateurs étant relativement peu renseignés.

[SLIDE 10]

La table de faits contient également nos 167 pays retenus, et plus de 3,600 indicateurs couvrant une période de 1970 à 2100, avec comme on l'a dit une proportion très importante de données manquantes ; on observe néanmoins un « pic » de données tous les 5 ans matérialisé par les flèches rouges sur le graphique de gauche. Nous avons donc choisi dans un premier temps de retenir les années de ces pics pour notre analyse historique, ce qui par ailleurs est cohérent avec les données prévisionnelles présentes dans cette table qui sont fournies tous les 5 ans entre 2020 et 2100.

[SLIDE 11]

La table des notes méthodologiques quant à elle, si elle contient bien également les 167 pays retenus, présente une grande variabilité des informations disponibles pour chaque année, matérialisée ici par le graphique en bas à gauche, sur lequel on voit 2 « pics » d'information en 2003 et 2012, avec une information quasi-inexistante à partir de 2017. Cette table ne couvre par ailleurs qu'un peu plus de 1,500 indicateurs sur les plus de 3,600 présents dans la table de faits, et aucun des indicateurs prospectifs que nous avons retenus n'y est représenté.

[SLIDE 12]

Enfin concernant la table de description et classement des indicateurs par thèmes, on note ici encore une forte proportion de valeurs manquantes mais aussi une très grande inégalité de répartition des indicateurs par thème, visible ici sur le graphique de droite, allant d'1 seul indicateur pour certains thèmes à plusieurs centaines voire plus de 1,000 pour d'autres.

[SLIDE 13]

Nous avons donc en premier lieu choisi uniquement les thèmes pertinents pour notre analyse, au nombre de 9 sur les 37 thèmes disponibles au total, à savoir : PIB, infrastructure et télécommunications, accès à l'enseignement secondaire et supérieur, à l'éducation et population de 15 à 24 ans, alphabétisation et population totale. Nous avons ensuite parcouru la liste des définitions de ces indicateurs pour sélectionner les plus pertinents, parfois en utilisant des mots-clés pour affiner notre sélection par tranche d'âges. Il faut noter que l'absence d'une description de la méthodologie de nommage des indicateurs ne nous a pas permis d'utiliser une méthode plus efficace de sélection, comme les RegEx par exemple. Cette sélection nous a conduits à retenir 13 indicateurs historiques et 7 indicateurs prospectifs, couvrant donc 10 thèmes au total.

[SLIDE 14]

La table sur cette diapositive résume les définitions des indicateurs historiques retenus, et est complétée comme on l'a dit par …

[SLIDE 15]

… les 4 indicateurs sélectionnés précédemment à la section 1.2 et qui couvrent, eux, les thèmes du PNB et du PIB par habitant ainsi que la population totale et son taux de croissance. En pratique, nos analyses graphiques exploratoires révéleront un taux de remplissage très inégal pour ces indicateurs, qui nous obligera à en éliminer certains au fur et à mesure de la progression de notre analyse, et nous passerons donc progressivement de 17 à 5 indicateurs historiques pertinents et suffisamment denses en information pour être utilisables dans notre processus de décision.

[SLIDE 16]

L'analyse préalable de ces indicateurs sur des radar charts fait apparaître l'Asie Orientale ainsi que l'Europe et l'Asie Centrale comme des régions d'intérêt particulier sur la période 2000-2015, suivies par l'Amérique Latine, qui comble son écart de façon

notable sur cette période du point de vue des indicateurs retenus ici, qui sont le nombre de PC pour 100 habitants, le taux d'accès à internet, les taux de scolarisation dans l'enseignement secondaire et universitaire et le taux d'alphabétisation. On voit bien ici comme la disponibilité des données est un problème pour certains indicateurs, avec le nombre de PC pour 100 habitants qui n'existe plus après 2010 et qui se matérialise par une ligne droite sur le radar chart de l'année 2015. Un focus sur l'année 2015 sur le graphique du milieu en bas montre que l'Asie Orientale et Méridionale ont les plus fortes populations totales, et que la croissance de la population la plus forte est observée sur la région Middle East & North Africa.

[SLIDE 17]

Concernant à présent les indicateurs prospectifs, nous en avons retenus 7, qui couvrent le thème du niveau scolaire le plus haut atteint par tranche d'âge.

[SLIDE 18]

L'analyse graphique de ces indicateurs sur la période 2020-2100 met en évidence plusieurs tendances, en particulier une tendance de long terme à la forte décroissance de la population qui ne finit pas le lycée chez les 15-24 ans de toutes régions, sauf en Afrique Sub-Saharienne où l'illettrisme augmente fortement pendant cette période. La population 15-24 ans finissant le lycée reste relativement stable dans toutes les régions, sauf en Asie Méridionale (en jaune ici en haut) où elle augmente fortement avant de se stabiliser, et en Asie Orientale et Pacific (en orange en haut), où elle décline constamment.

[SLIDE 19]

Sur la même période de 2020 à 2100, on note un recul de l'illettrisme très fort en Asie Méridionale (en jaune sur les graphiques du haut) et en Afrique Sub-Saharienne (en beige sur ces mêmes graphiques), tandis que l'augmentation de la population ayant un niveau d'éducation universitaire (représentée sur le graphique du bas) augmente fortement dans ces mêmes régions. Il conviendrait de vérifier par ailleurs ces chiffres,

l’augmentation relativement forte du nombre d’étudiants ayant une éducation universitaire en Afrique Sub-Saharienne semblant *a priori* contradictoire avec le fait que sur la même période la population de 15-24 qui ne finit pas le lycée est également en forte augmentation pour cette région. La table de données méthodologiques ne contenant pas d’informations suffisantes à cet égard, nous n’avons pas pu élucider cette contradiction apparente.

[SLIDE 20]

Afin de décider quels pays cibler pour notre expansion internationale, nous avons choisi d’établir un classement de notre échantillon de pays en fonction de 6 critères historiques, à savoir : le taux d’accès à internet, le taux d’inscription dans l’enseignement secondaire et universitaire, la population de 15 à 24 ans, le taux de croissance moyen de la population et le PIB par habitant. Afin d’améliorer la qualité des résultats, il convient donc de retraiter les valeurs manquantes. Les supprimer reviendrait à éliminer trop de pays de notre échantillon, donc nous avons choisi de procéder à des imputations.

[SLIDE 21]

Avant d'appliquer un ranking aux différents pays sur chacun des indicateurs, nous avons décidé de procéder au remplissage des valeurs manquantes par la valeur médiane pour cet indicateur pour le groupe de revenus auquel le pays appartient. Notre cible principale étant les pays à haut revenu et à revenu moyen supérieur, le niveau de développement de ces pays porte à penser que les valeurs manquantes y sont très certainement dues à un problème de collecte et/ou de retraitement, plutôt qu'a l'absence de ces indicateurs ou à leur nullité, et l'imputation par la médiane permet donc de ne pas défavoriser ces pays intéressants pour notre stratégie pour un manque relatif de certaines données. L'imputation par la médiane apparaît comme une mesure raisonnable dans notre cas en raison de la présence d'outliers dans les groupes de revenus au niveau de certains indicateurs, et de la distribution relativement aléatoire des valeurs manquantes, et constitue une des imputations les plus neutres au vu des données disponibles, permettant comme on l'a dit de ne pas défavoriser certains pays tout en n'introduisant pas de biais dans l'analyse des rangs.

[SLIDE 22]

Cette décision est par ailleurs également soutenue par la présence d'outliers pour presque tous les indicateurs et les groupes de revenus, matérialisés ici par les cercles rouges au-dessus ou en dessous des moustaches de nos boxplots. Au vu des valeurs manquantes et des outliers par groupe de revenus, la médiane est donc un paramètre satisfaisant pour l'imputation des valeurs manquantes par pays, car elle n'est pas affectée par la présence d'outliers ni de valeurs manquantes.

[SLIDE 23]

La méthodologie de classement que nous avons appliquée a donc consisté à calculer le dense rank (sans *ex-æquo*) de chaque pays pour chacun des 6 indicateurs retenus pour l'année 2015, puis de calculer le rang moyen des pays avant de les classer et de regarder les 20 premiers. Nous avons également calculé un percentile rank pour chaque pays et chaque indicateur, pour évaluer l'importance relative de chacun des 6 indicateurs sous-jacents dans le classement final des pays.

[SLIDE 24]

Cette méthodologie nous a permis d'arriver au classement suivant, où 5 pays parmi les 20 premiers sont anglophones, et 3 sont francophones. On voit que les pays à hauts revenus dominent largement le classement avec 19 pays sur 20, confirmant notre analyse initiale, ainsi que les pays d'Europe & Asie centrale (avec 10 pays sur 20), suivis par l'Asie Orientale et le Pacifique.

[SLIDE 25]

En termes de stratégie, il conviendrait en 2024 de se concentrer sur une étude de compétitivité-prix et une étude des barrières à l'entrée sur les marchés hors-Europe avant en 2025 de commencer l'expansion vers les pays anglophones du top 5 (États-Unis, Grande-Bretagne et Australie), tout en commençant à étudier la faisabilité de réorientation partielle vers le marché de la formation professionnelle & diplômante en

France (financements CPF). D'ici à fin 2027, l'accent devrait être mis sur les pays anglophones du top 20 (top 5 + Canada, Nouvelle-Zélande & Irlande) ainsi que tactiquement sur d'autres marchés à forte valeur ajoutée comme l'Arabie Saoudite, Oman et le Koweït, avant éventuellement d'envisager d'ici à 2030 une diversification plus poussée en Europe vers les pays non-anglophones et non-francophones de cette zone et d'envisager le redéploiement du catalogue français vers des formations professionnelles & diplômantes, qui seraient de fait également reconnues dans les autres pays européens en vertu de la reconnaissance des diplômes.

[SLIDE 26]

En conclusion, le jeu de données fourni peut effectivement, dans une certaine mesure, informer les décisions stratégiques d'expansion internationale de l'entreprise, mais il convient d'être prudent et de ne…

- …pas l'utiliser en première analyse :
  - Il faut évaluer le positionnement concurrentiel prix-produit pour déterminer la stratégie en amont
  - Et penser la gamme de produits dans un contexte international : où sont les besoins et quels sont-ils ? La gamme de cours en ligne actuelle y est-elle adaptée ?
  - Il faut aussi repenser le marché domestique ;
  - Et évaluer les barrières à l'entrée réglementaires des différents marchés (l'éducation étant la prérogative des états, les conditions de faisabilité de l'expansion hors-EU seront à étudier avec précision) ;
- …et il ne faut pas l'utiliser en unique source de données chiffrées :
  - Il faudrait intégrer d'autres indicateurs comme les classements PISA et les pratiques éducatives extra-scolaires (cf. Singapour par exemple), ainsi que d'évaluer au mieux le champ concurrentiel domestique & international ;

- …il conviendra enfin de ne pas utiliser intégralement ce jeu de données, qui contient comme on l'q vu une multiplicité d'indicateurs inutiles et de faible densité informationnelle ; il faudra mieux cibler la collecte des données pour gagner en pertinence dans l'analyse.

Ce qui nous amène à la fin de cette présentation.